Case Study: Instacart

Domain: E-commerce

Instacart is a grocery ordering and delivery app that aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. Instacart’s data science team plays a big part in providing this

delightful shopping experience. Currently,they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session.

The dataset is a relational set of files describing customers' orders over time. The

dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.

Tasks:

As a Big Data consultant,you are helping the data science team to deal with

a large amount of data. To solve this, you are pitching in to move the transactional data from RDBMS to HDFS by:

1.Verify the cluster health including HDFS and Spark

2.Test the spark environment by executing the spark’s sort.py example.

3.Try to implement the same example in scala and perform spark-submit.

4.Analyze the behavior of spark application on Spark web UI

5.Add custom logs in your application and re-execute the application. Once

executed check the Spark logs.

6.Transfer complete dataset from RDBMS to HDFS

7.Validate the loaded data by comparing the statistics of data both in source

and HDFS

8.Create a new directory in HDFS named cheeses and load only rows where

aisle is “specialty cheeses“

9.update “specialty cheeses” to “specialtycheese” and transfer only updated

rows in the above-created directory.

1.Verify the cluster health including HDFS and Spark

[raghavendr48edu@ip-10-1-1-204 ~]$ **hdfs fsck /user/raghavendr48edu**

Connecting to namenode via  <http://ip-10-1-2-24.ap-south-1.compute.internal:9870/fsck?ugi=raghavendr48edu&path=%2Fuser%2Fraghavendr48edu>

FSCK started by raghavendr48edu (auth:SIMPLE) from /10.1.1.204 for path /user/raghavendr48edu at Sun May 21 15:41:05 UTC 2023

Status: HEALTHY

Number of data-nodes: 3

Number of racks: 1

Total dirs: 39

Total symlinks: 0

Replicated Blocks:

Total size: 1682851351 B

Total files: 93

Total blocks (validated): 82 (avg. block size 20522577 B)

Minimally replicated blocks: 82 (100.0 %)

Over-replicated blocks: 0 (0.0 %)

Under-replicated blocks: 0 (0.0 %)

Mis-replicated blocks: 0 (0.0 %)

Default replication factor: 3

Average block replication: 3.0

Missing blocks: 0

Corrupt blocks: 0

Missing replicas: 0 (0.0 %)

Blocks queued for replication: 0

Erasure Coded Block Groups:

Total size: 0 B

Total files: 0

Total block groups (validated): 0

Minimally erasure-coded block groups: 0

Over-erasure-coded block groups: 0

Under-erasure-coded block groups: 0

Unsatisfactory placement block groups: 0

Average block group size: 0.0

Missing block groups: 0

Corrupt block groups: 0

Missing internal blocks: 0

Blocks queued for replication: 0

FSCK ended at Sun May 21 15:41:05 UTC 2023 in 2 milliseconds

The filesystem under path '/user/raghavendr48edu' is HEALTHY

[raghavendr48edu@ip-10-1-1-204 ~]$

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

2.Test the spark environment by executing the spark’s sort.py example.

**sort.py**

from \_\_future\_\_ import print\_function  
import sys,time  
from pyspark.sql import SparkSession  
if \_\_name\_\_ == "\_\_main\_\_":  
 if len(sys.argv) != 2:  
 print("Usage: hdfstest.py <file>", file=sys.stderr)  
 exit(-1)  
 spark = SparkSession\  
 .builder\  
 .appName("HdfsTest\_py")\  
 .getOrCreate()  
 file\_ = spark.read.text(sys.argv[1]).rdd  
 mapped = file\_.map(lambda s:len(s)).cache()  
 for i in range(10):  
 start\_time = time.time()  
 mapped.map(lambda x: x+2)  
 end\_time = time.time()  
 print('----> Iteration took:', end\_time - start\_time,'ms')  
 spark.stop()

[raghavendr48edu@ip-10-1-1-204 ~]$ **spark-submit --class org.apache.spark.examples.HdfsTest --master yarn --deploy-mode cluster /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/examples/jar**

**s/spark-examples\*.jar use\_cases/sort.py**

23/05/23 14:47:19 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/23 14:47:19 INFO yarn.Client: Requesting a new application from cluster with 3 NodeManagers

23/05/23 14:47:19 INFO conf.Configuration: resource-types.xml not found

23/05/23 14:47:19 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/23 14:47:19 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (20480 MB per container)

23/05/23 14:47:19 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead

23/05/23 14:47:19 INFO yarn.Client: Setting up container launch context for our AM

23/05/23 14:47:19 INFO yarn.Client: Setting up the launch environment for our AM container

23/05/23 14:47:19 INFO yarn.Client: Preparing resources for our AM container

23/05/23 14:47:19 INFO yarn.Client: Uploading resource [file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/examples/jars/spark-examples\_2.11-2.4.0-cdh6.2.1.jar](file:///\\opt\cloudera\parcels\CDH-6.2.1-1.cdh6.2.1.p0.1425774\lib\spark\examples\jars\spark-examples_2.11-2.4.0-cdh6.2.1.jar) -> hdfs://nameservice1/user/ragh

avendr48edu/.sparkStaging/application\_1684298513961\_1255/spark-examples\_2.11-2.4.0-cdh6.2.1.jar

23/05/23 14:47:20 INFO yarn.Client: Uploading resource [file:/tmp/spark-f5c1bfcd-cc65-4415-9d44-f40bcc336dfa/\_\_spark\_conf\_\_7880736034069233479.zip](file:///\\tmp\spark-f5c1bfcd-cc65-4415-9d44-f40bcc336dfa\__spark_conf__7880736034069233479.zip) -> hdfs://nameservice1/user/raghavendr48edu/.sparkStaging/appli

cation\_1684298513961\_1255/\_\_spark\_conf\_\_.zip

23/05/23 14:47:20 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/23 14:47:20 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/23 14:47:20 INFO spark.SecurityManager: Changing view acls groups to:

23/05/23 14:47:20 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/23 14:47:20 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu); groups with view permissions: Set(); users with mo

dify permissions: Set(raghavendr48edu); groups with modify permissions: Set()

23/05/23 14:47:20 INFO conf.HiveConf: Found configuration file [file:/etc/hive/conf.cloudera.hive/hive-site.xml](file:///\\etc\hive\conf.cloudera.hive\hive-site.xml)

23/05/23 14:47:20 INFO security.YARNHadoopDelegationTokenManager: Attempting to load user's ticket cache.

23/05/23 14:47:20 INFO yarn.Client: Submitting application application\_1684298513961\_1255 to ResourceManager

23/05/23 14:47:20 INFO impl.YarnClientImpl: Submitted application application\_1684298513961\_1255

23/05/23 14:47:21 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: ACCEPTED)

23/05/23 14:47:21 INFO yarn.Client:

client token: N/A

diagnostics: AM container is launched, waiting for AM container to Register with RM

ApplicationMaster host: N/A

ApplicationMaster RPC port: -1

queue: root.default

start time: 1684853240416

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_1255/>

user: raghavendr48edu

23/05/23 14:47:22 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: ACCEPTED)

23/05/23 14:47:23 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: ACCEPTED)

23/05/23 14:47:24 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: ACCEPTED)

23/05/23 14:47:25 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:25 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: ip-10-1-2-103.ap-south-1.compute.internal

ApplicationMaster RPC port: 39437

queue: root.default

start time: 1684853240416

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_1255/>

user: raghavendr48edu

23/05/23 14:47:26 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:27 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:28 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:29 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:30 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:31 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:32 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:33 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:34 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:35 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:36 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:37 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:38 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:39 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:40 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:41 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:42 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:43 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:44 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:45 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:46 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:47 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:48 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:49 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:50 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:51 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:52 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:53 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:54 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:55 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:56 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:57 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: RUNNING)

23/05/23 14:47:58 INFO yarn.Client: Application report for application\_1684298513961\_1255 (state: FINISHED)

23/05/23 14:47:58 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: ip-10-1-2-103.ap-south-1.compute.internal

ApplicationMaster RPC port: 39437

queue: root.default

start time: 1684853240416

final status: SUCCEEDED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_1255/>

user: raghavendr48edu

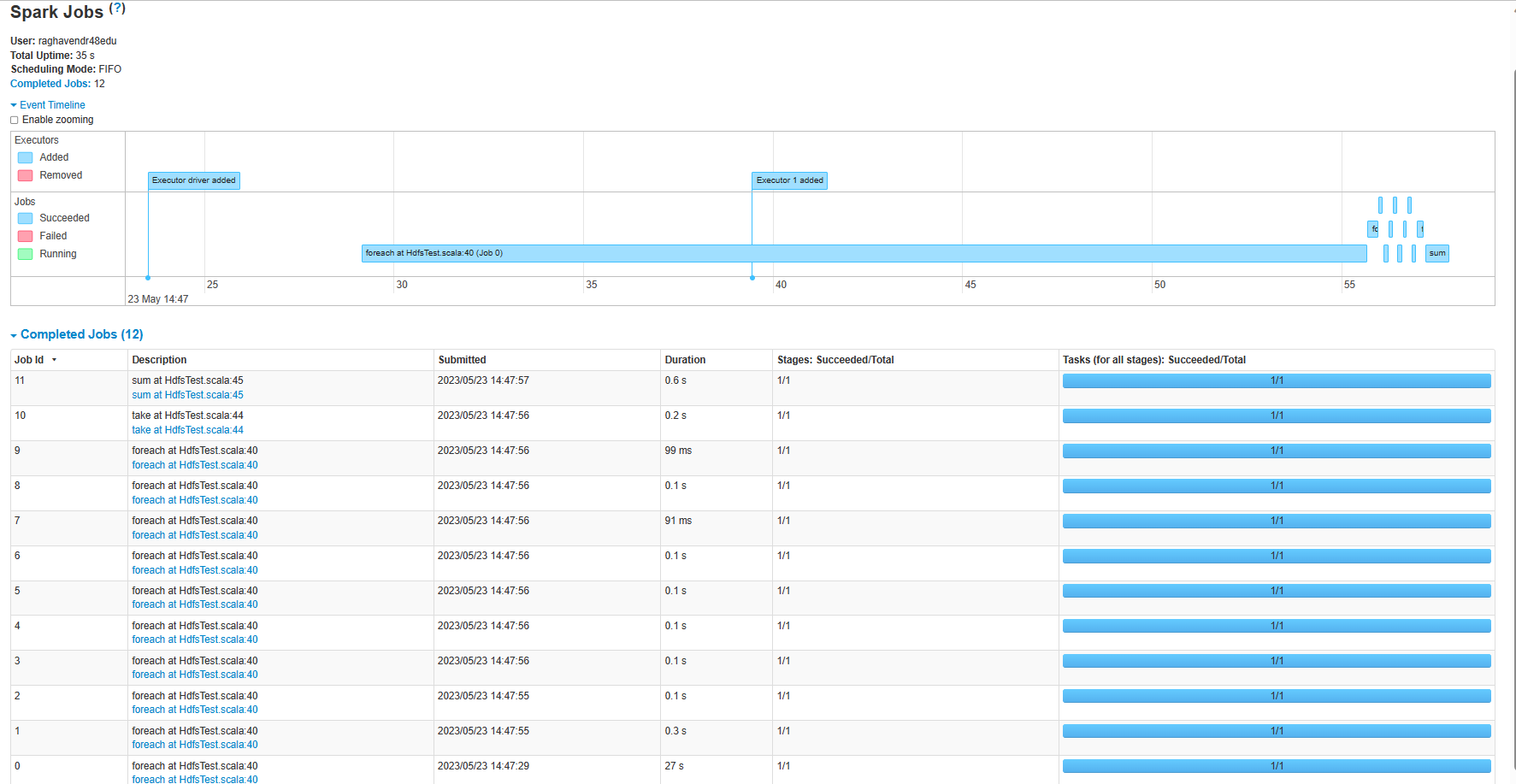
23/05/23 14:47:58 INFO util.ShutdownHookManager: Shutdown hook called

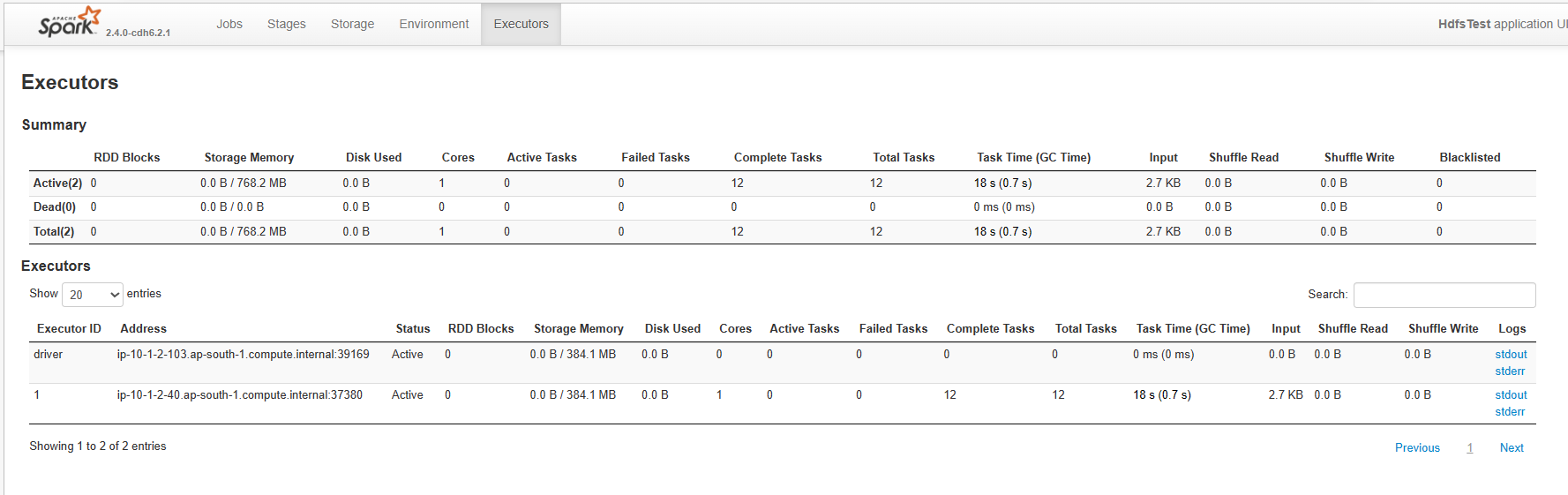
23/05/23 14:47:58 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-95d02cc4-20da-4312-913f-9682275dd16a

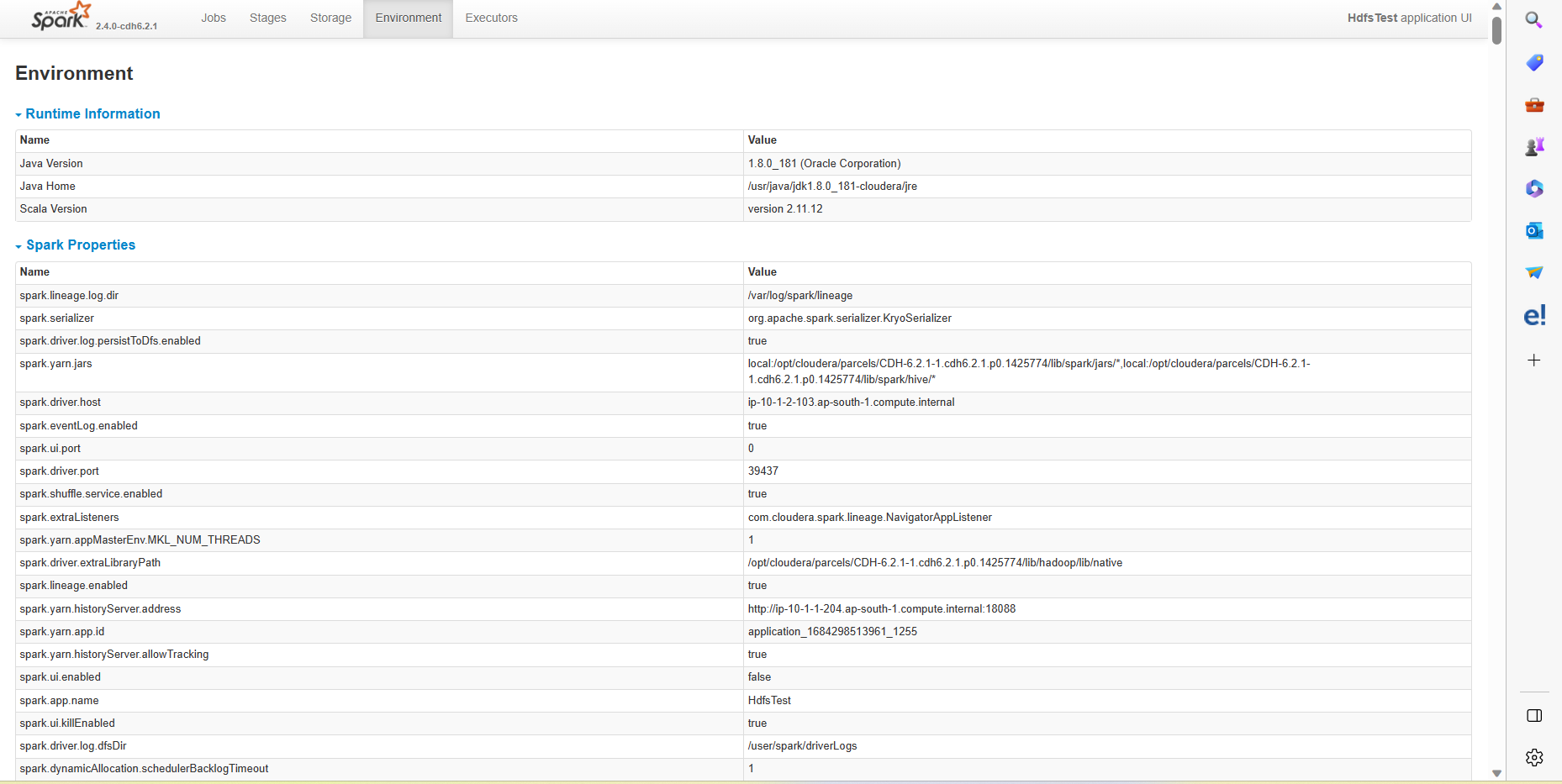
23/05/23 14:47:58 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-f5c1bfcd-cc65-4415-9d44-f40bcc336dfa

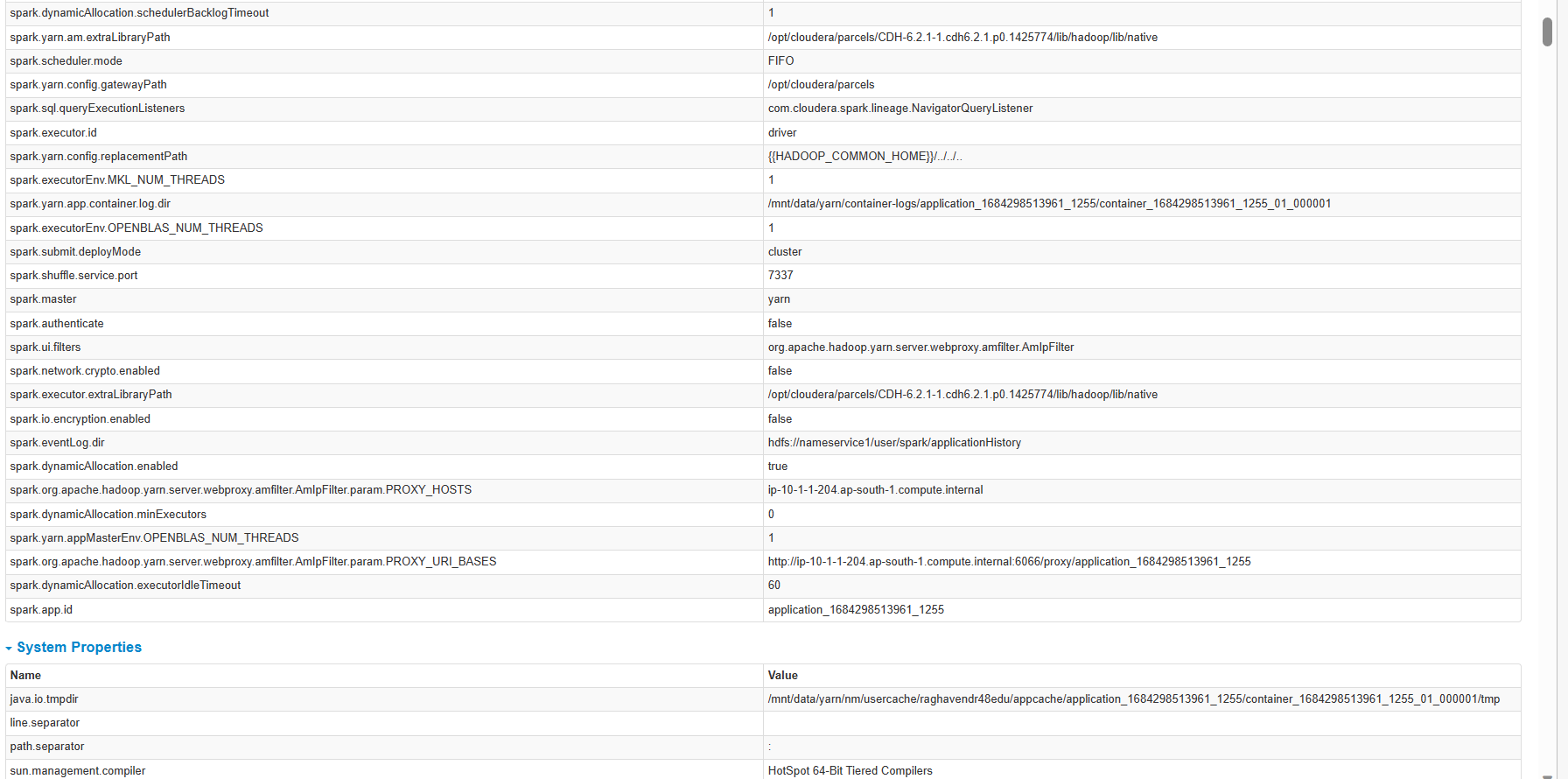
*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

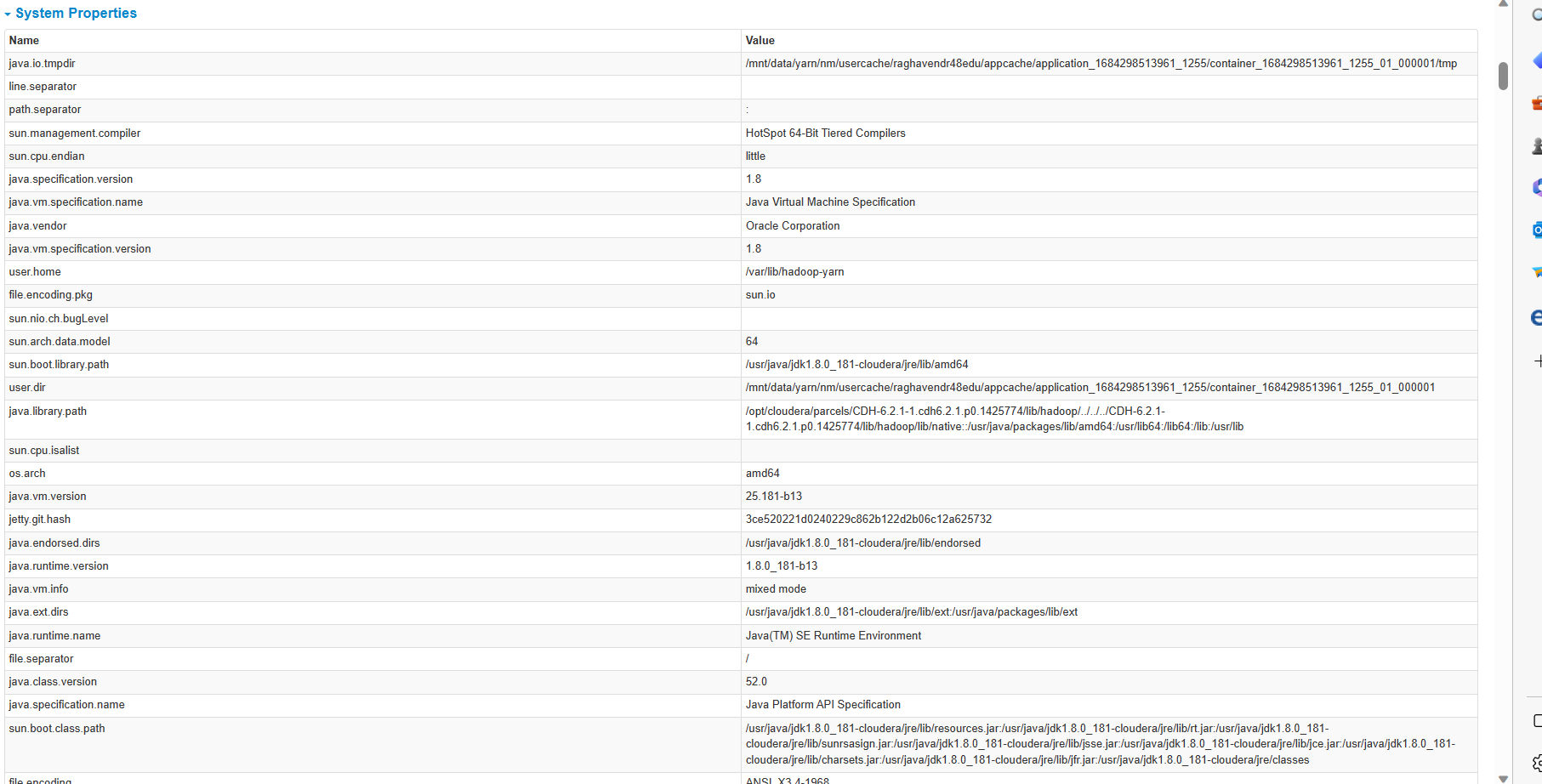
4.Analyze the behavior of spark application on Spark web UI

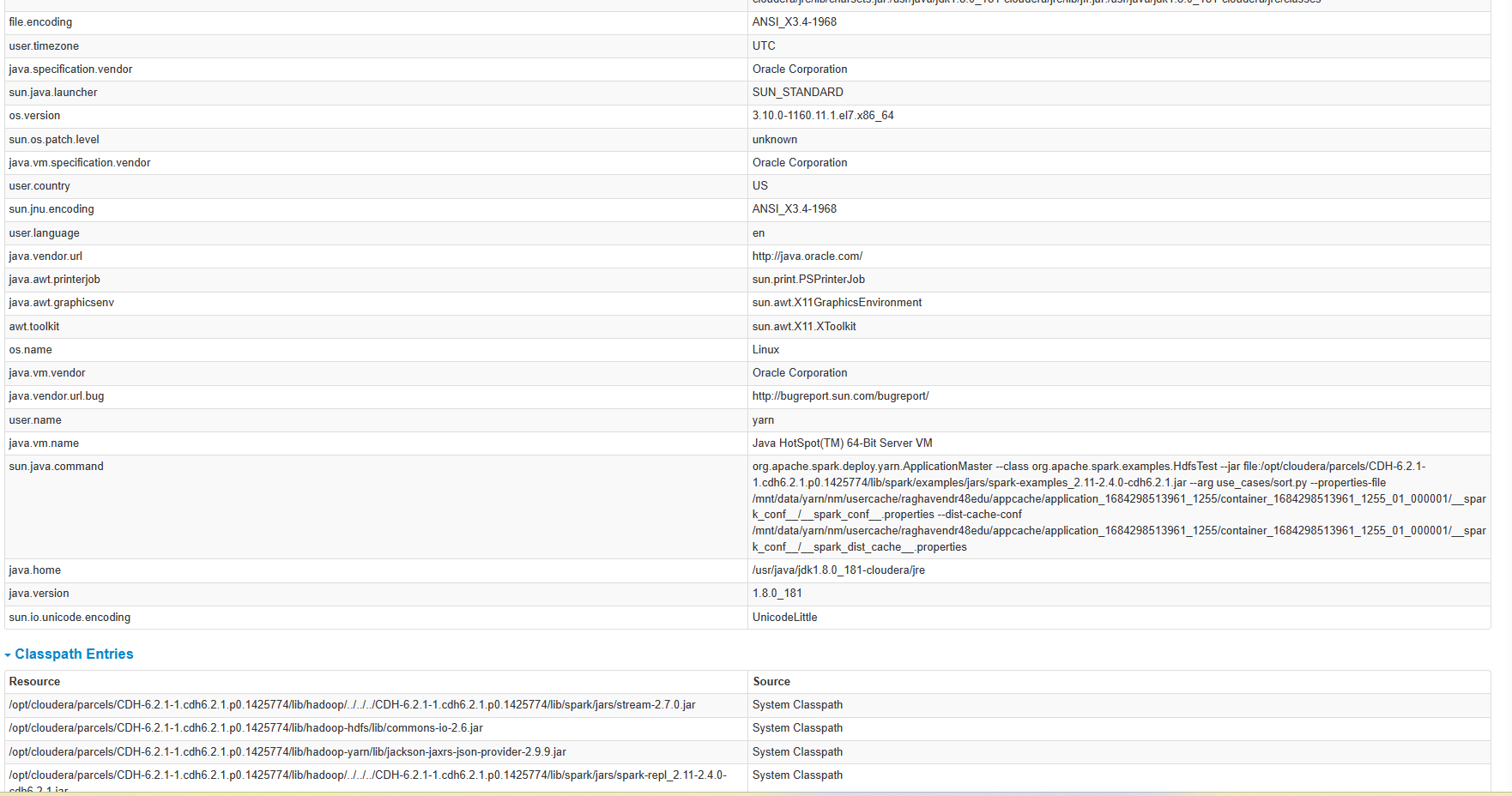


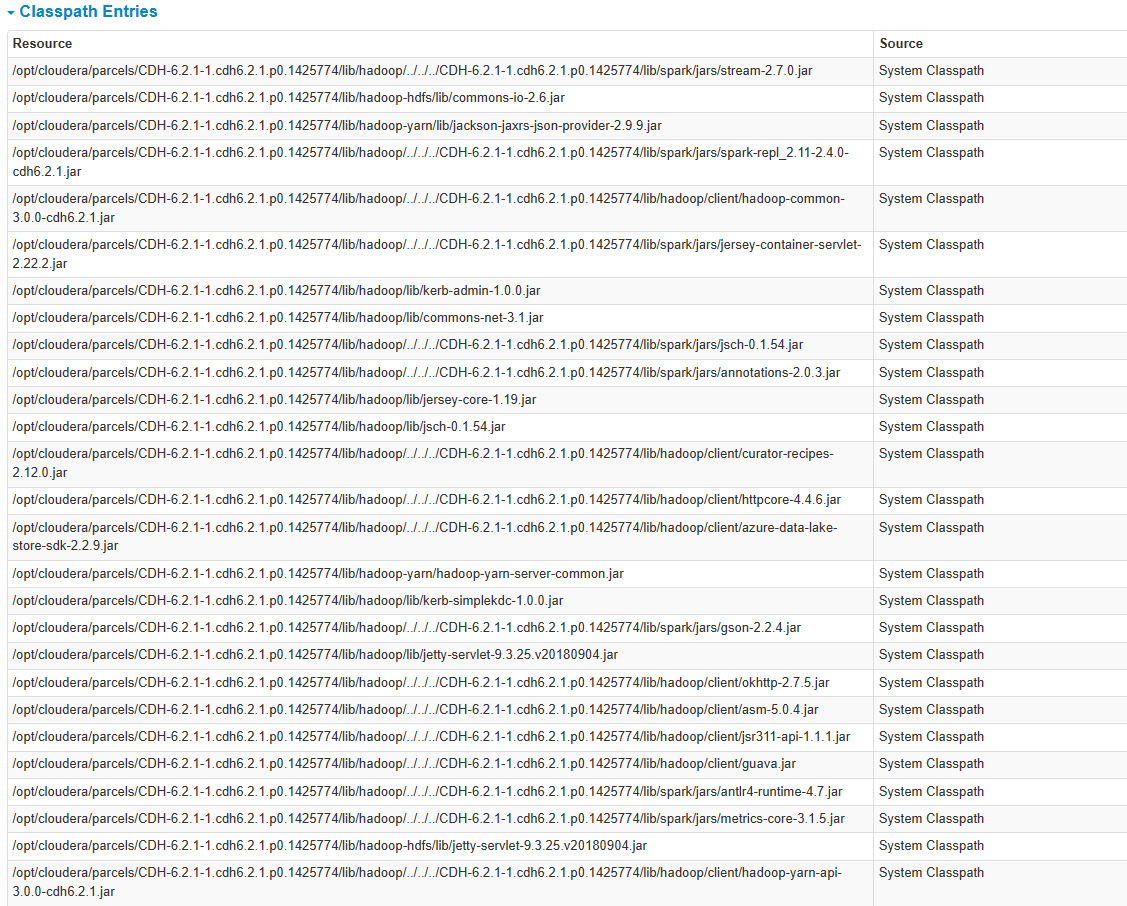












OR

[raghavendr48edu@ip-10-1-1-204 ~]$ **hdfs dfs -copyFromLocal /home/raghavendr48edu/input\_sort\_py.txt**

[raghavendr48edu@ip-10-1-1-204 ~]$ **hdfs dfs -copyFromLocal /home/raghavendr48edu/sort.py**

[raghavendr48edu@ip-10-1-1-204 ~]$ **spark-submit sort.py input\_sort\_py.txt**

23/05/25 14:09:41 INFO spark.SparkContext: Running Spark version 2.4.0-cdh6.2.1

23/05/25 14:09:41 INFO logging.DriverLogger: Added a local log appender at: /tmp/spark-25931366-28b8-4833-8844-e4784f9f2db5/\_\_driver\_logs\_\_/driver.log

23/05/25 14:09:41 INFO spark.SparkContext: Submitted application: HdfsTest\_py

23/05/25 14:09:41 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/25 14:09:41 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/25 14:09:41 INFO spark.SecurityManager: Changing view acls groups to:

23/05/25 14:09:41 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/25 14:09:41 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu); groups with view permissions: Set(); users with modify permissions: Set(ragh

avendr48edu); groups with modify permissions: Set()

23/05/25 14:09:41 INFO util.Utils: Successfully started service 'sparkDriver' on port 39223.

23/05/25 14:09:41 INFO spark.SparkEnv: Registering MapOutputTracker

23/05/25 14:09:41 INFO spark.SparkEnv: Registering BlockManagerMaster

23/05/25 14:09:41 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information

23/05/25 14:09:41 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up

23/05/25 14:09:41 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-4066e0bb-e134-4674-9a27-d12a24906ec5

23/05/25 14:09:41 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB

23/05/25 14:09:41 INFO spark.SparkEnv: Registering OutputCommitCoordinator

23/05/25 14:09:41 INFO util.Utils: Using initial executors = 0, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances

23/05/25 14:09:42 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/25 14:09:42 INFO yarn.Client: Requesting a new application from cluster with 3 NodeManagers

23/05/25 14:09:42 INFO conf.Configuration: resource-types.xml not found

23/05/25 14:09:42 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/25 14:09:42 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (20480 MB per container)

23/05/25 14:09:42 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead

23/05/25 14:09:42 INFO yarn.Client: Setting up container launch context for our AM

23/05/25 14:09:42 INFO yarn.Client: Setting up the launch environment for our AM container

23/05/25 14:09:42 INFO yarn.Client: Preparing resources for our AM container

23/05/25 14:09:43 INFO yarn.Client: Uploading resource [file:/tmp/spark-25931366-28b8-4833-8844-e4784f9f2db5/\_\_spark\_conf\_\_6655970108633587660.zip](file:///\\tmp\spark-25931366-28b8-4833-8844-e4784f9f2db5\__spark_conf__6655970108633587660.zip) -> hdfs://nameservice1/user/raghavendr48edu/.sparkStaging/application\_1684866872278\_1298/

\_\_spark\_conf\_\_.zip

23/05/25 14:09:43 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/25 14:09:43 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/25 14:09:43 INFO spark.SecurityManager: Changing view acls groups to:

23/05/25 14:09:43 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/25 14:09:43 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu); groups with view permissions: Set(); users with modify permissions: Set(ragh

avendr48edu); groups with modify permissions: Set()

23/05/25 14:09:43 INFO yarn.Client: Submitting application application\_1684866872278\_1298 to ResourceManager

23/05/25 14:09:43 INFO impl.YarnClientImpl: Submitted application application\_1684866872278\_1298

23/05/25 14:09:44 INFO yarn.Client: Application report for application\_1684866872278\_1298 (state: ACCEPTED)

23/05/25 14:09:44 INFO yarn.Client:

client token: N/A

diagnostics: AM container is launched, waiting for AM container to Register with RM

ApplicationMaster host: N/A

ApplicationMaster RPC port: -1

queue: root.default

start time: 1685023783424

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_1298/>

user: raghavendr48edu

23/05/25 14:09:45 INFO yarn.Client: Application report for application\_1684866872278\_1298 (state: ACCEPTED)

23/05/25 14:09:46 INFO cluster.YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY\_HOSTS -> ip-10-1-1-204.ap-south-1.compute.internal, PROXY\_URI\_BASES -> [http://ip-10-1-1-](http://ip-10-1-1-/)

204.ap-south-1.compute.internal:6066/proxy/application\_1684866872278\_1298), /proxy/application\_1684866872278\_1298

23/05/25 14:09:46 INFO yarn.Client: Application report for application\_1684866872278\_1298 (state: RUNNING)

23/05/25 14:09:46 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: 10.1.2.103

ApplicationMaster RPC port: -1

queue: root.default

start time: 1685023783424

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_1298/>

user: raghavendr48edu

23/05/25 14:09:46 INFO cluster.YarnClientSchedulerBackend: Application application\_1684866872278\_1298 has started running.

23/05/25 14:09:46 INFO cluster.SchedulerExtensionServices: Starting Yarn extension services with app application\_1684866872278\_1298 and attemptId None

23/05/25 14:09:46 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 34041.

23/05/25 14:09:46 INFO netty.NettyBlockTransferService: Server created on ip-10-1-1-204.ap-south-1.compute.internal:34041

23/05/25 14:09:46 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy

23/05/25 14:09:46 INFO storage.BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 34041, None)

23/05/25 14:09:46 INFO storage.BlockManagerMasterEndpoint: Registering block manager ip-10-1-1-204.ap-south-1.compute.internal:34041 with 366.3 MB RAM, BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 34041, None)

23/05/25 14:09:46 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 34041, None)

23/05/25 14:09:46 INFO storage.BlockManager: external shuffle service port = 7337

23/05/25 14:09:46 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 34041, None)

23/05/25 14:09:46 INFO util.log: Logging initialized @7100ms

23/05/25 14:09:46 INFO scheduler.EventLoggingListener: Logging events to hdfs://nameservice1/user/spark/applicationHistory/application\_1684866872278\_1298

23/05/25 14:09:46 INFO util.Utils: Using initial executors = 0, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances

23/05/25 14:09:46 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!

23/05/25 14:09:46 INFO spark.SparkContext: Registered listener com.cloudera.spark.lineage.NavigatorAppListener

23/05/25 14:09:47 INFO logging.DriverLogger$DfsAsyncWriter: Started driver log file sync to: /user/spark/driverLogs/application\_1684866872278\_1298\_driver.log

23/05/25 14:09:47 INFO cluster.YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.8

23/05/25 14:09:47 INFO cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)

23/05/25 14:09:47 INFO internal.SharedState: loading hive config file: [file:/etc/hive/conf.cloudera.hive/hive-site.xml](file:///\\etc\hive\conf.cloudera.hive\hive-site.xml)

23/05/25 14:09:47 INFO internal.SharedState: spark.sql.warehouse.dir is not set, but hive.metastore.warehouse.dir is set. Setting spark.sql.warehouse.dir to the value of hive.metastore.warehouse.dir ('/user/hive/warehouse').

23/05/25 14:09:47 INFO internal.SharedState: Warehouse path is '/user/hive/warehouse'.

23/05/25 14:09:47 INFO state.StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint

23/05/25 14:09:49 INFO datasources.FileSourceStrategy: Pruning directories with:

23/05/25 14:09:49 INFO datasources.FileSourceStrategy: Post-Scan Filters:

23/05/25 14:09:49 INFO datasources.FileSourceStrategy: Output Data Schema: struct<value: string>

23/05/25 14:09:49 INFO execution.FileSourceScanExec: Pushed Filters:

23/05/25 14:09:49 INFO codegen.CodeGenerator: Code generated in 191.667246 ms

23/05/25 14:09:49 INFO memory.MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 337.3 KB, free 366.0 MB)

23/05/25 14:09:50 INFO memory.MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 32.3 KB, free 365.9 MB)

23/05/25 14:09:50 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:34041 (size: 32.3 KB, free: 366.3 MB)

23/05/25 14:09:50 INFO spark.SparkContext: Created broadcast 0 from javaToPython at NativeMethodAccessorImpl.java:0

23/05/25 14:09:50 INFO execution.FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes.

**----> Iteration took: 2.1696090698242188e-05 ms**

**----> Iteration took: 5.7220458984375e-06 ms**

**----> Iteration took: 4.0531158447265625e-06 ms**

**----> Iteration took: 3.814697265625e-06 ms**

**----> Iteration took: 4.0531158447265625e-06 ms**

**----> Iteration took: 3.337860107421875e-06 ms**

**----> Iteration took: 3.5762786865234375e-06 ms**

**----> Iteration took: 3.5762786865234375e-06 ms**

**----> Iteration took: 3.5762786865234375e-06 ms**

**----> Iteration took: 3.5762786865234375e-06 ms**

23/05/25 14:09:50 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread

23/05/25 14:09:50 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors

23/05/25 14:09:50 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down

23/05/25 14:09:50 INFO cluster.SchedulerExtensionServices: Stopping SchedulerExtensionServices

(serviceOption=None,

services=List(),

started=false)

23/05/25 14:09:50 INFO cluster.YarnClientSchedulerBackend: Stopped

23/05/25 14:09:50 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!

23/05/25 14:09:50 INFO memory.MemoryStore: MemoryStore cleared

23/05/25 14:09:50 INFO storage.BlockManager: BlockManager stopped

23/05/25 14:09:50 INFO storage.BlockManagerMaster: BlockManagerMaster stopped

23/05/25 14:09:50 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

23/05/25 14:09:50 INFO spark.SparkContext: Successfully stopped SparkContext

23/05/25 14:09:51 INFO util.ShutdownHookManager: Shutdown hook called

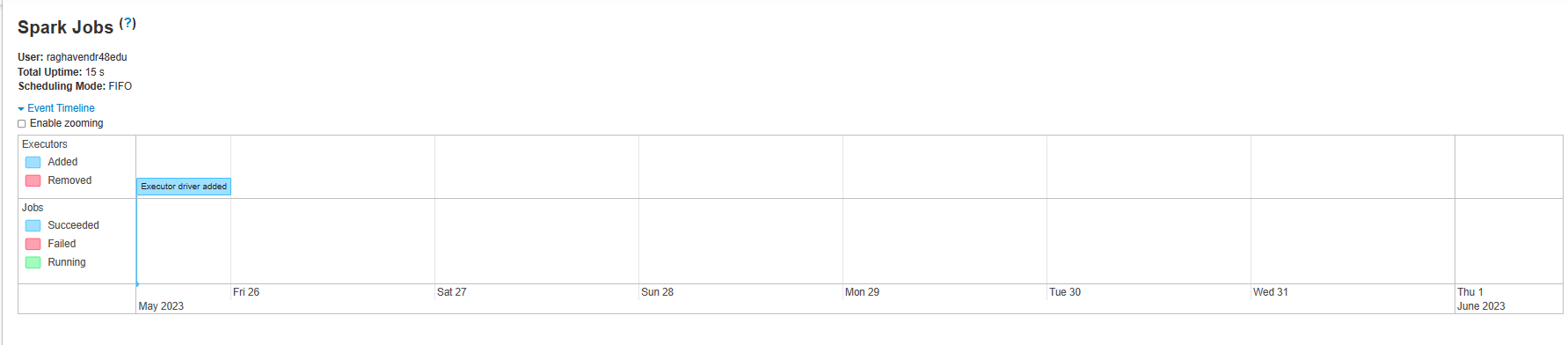
23/05/25 14:09:51 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-25931366-28b8-4833-8844-e4784f9f2db5

23/05/25 14:09:51 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-a9c865b8-3b7f-4df6-af17-e8dcf7942b93

23/05/25 14:09:51 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-25931366-28b8-4833-8844-e4784f9f2db5/pyspark-fa8807c1-85ab-4fae-bf1e-c0201f7e0139

[raghavendr48edu@ip-10-1-1-204 ~]$

4.Analyze the behavior of spark application on Spark web UI



**Changed sort.py code and reexecuted spark-submit sort.py input\_sort\_py.txt**

from \_\_future\_\_ import print\_function

import sys,time

from pyspark.sql import SparkSession

if \_\_name\_\_ == "\_\_main\_\_":

if len(sys.argv) != 2:

print("Usage: hdfstest.py <file>", file=sys.stderr)

exit(-1)

spark = SparkSession\

.builder\

.appName("HdfsTest\_py")\

.getOrCreate()

file\_ = spark.read.text(sys.argv[1]).rdd

mapped = file\_.map(lambda s:len(s)).cache()

for i in range(10):

start\_time = time.time()

mapped.map(lambda x: x+2)

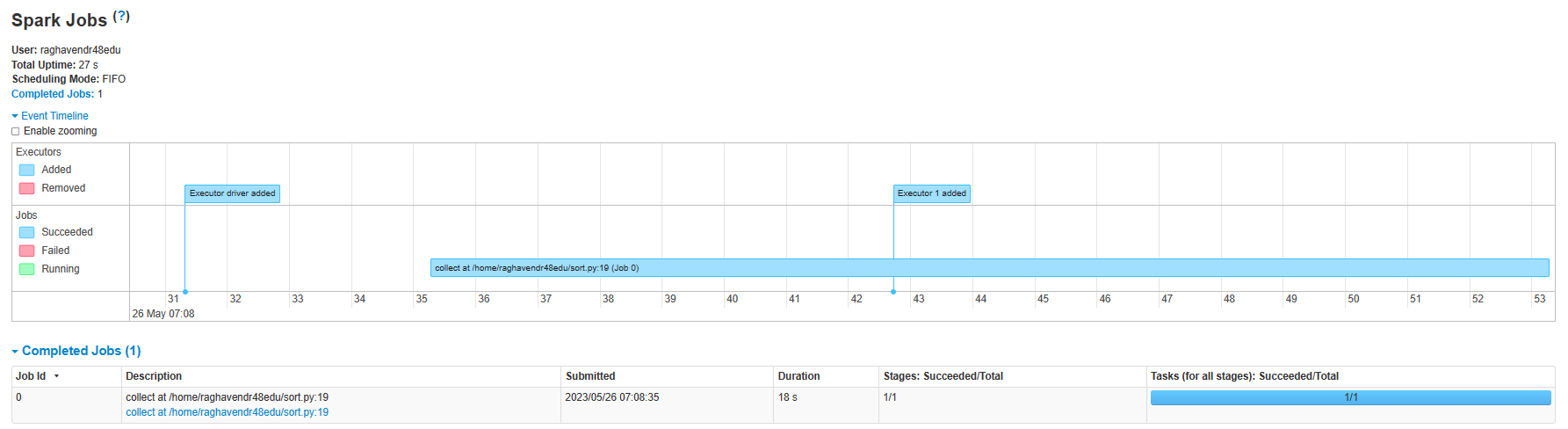
end\_time = time.time()

print('----> Iteration took:', end\_time - start\_time,'ms')

result=mapped.collect()

print(result)

spark.stop()



3.Try to implement the same example in scala and perform spark-submit.

**Step 1: Create a folder named training\_project mkdir training\_project**

**Step 2: Create build.sbt in the training\_project folder with the following content (using vi build.sbt)**

name := "SparkMe Project"

version := "1.0"

organization := "edureka"

scalaVersion := "2.11.8"

val sparkVersion = "2.1.0"

libraryDependencies += "org.apache.spark" %% "spark-core" % sparkVersion %"provided"

libraryDependencies += "org.apache.spark" %% "spark-sql" % sparkVersion % "provided"

resolvers += Resolver.mavenLocal



**Step 3: Verify build.sbt by executing the command given below sbt compile**

sbt compile

The above command will create directory “project” and “target”.

[raghavendr48edu@ip-10-1-1-204 training\_project3]$ **sbt compile**

[info] welcome to sbt 1.5.5 (Oracle Corporation Java 1.8.0\_181)

[info] loading project definition from /home/raghavendr48edu/training\_project3/project

[info] loading settings for project training\_project3 from build.sbt ...

[info] set current project to SparkMe Project (in build [file:/home/raghavendr48edu/training\_project3/](file:///\\home\raghavendr48edu\training_project3\))

[info] Executing in batch mode. For better performance use sbt's shell

[success] Total time: 1 s, completed May 25, 2023 2:16:27 PM

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

[raghavendr48edu@ip-10-1-1-204 training\_project3]$ **spark-submit --class HdfsTest --deploy-mode client target/scala-2.11/sparkme-project\_2.11-1.0.jar input\_sort\_py.txt**

23/05/25 12:27:09 INFO spark.SparkContext: Running Spark version 2.4.0-cdh6.2.1

23/05/25 12:27:09 INFO logging.DriverLogger: Added a local log appender at: /tmp/spark-149fb59c-5b64-4b50-a117-6212d417e2c9/\_\_driver\_logs\_\_/driver.log

23/05/25 12:27:09 INFO spark.SparkContext: Submitted application: CustomHdfsTest

23/05/25 12:27:09 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/25 12:27:09 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/25 12:27:09 INFO spark.SecurityManager: Changing view acls groups to:

23/05/25 12:27:09 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/25 12:27:09 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu);

groups with view permissions: Set(); users with modify permissions: Set(raghavendr48edu); groups with modify permissions: Set()

23/05/25 12:27:09 INFO util.Utils: Successfully started service 'sparkDriver' on port 40066.

23/05/25 12:27:09 INFO spark.SparkEnv: Registering MapOutputTracker

23/05/25 12:27:09 INFO spark.SparkEnv: Registering BlockManagerMaster

23/05/25 12:27:09 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information

23/05/25 12:27:09 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up

23/05/25 12:27:09 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-7f3f18dd-6ed8-4d58-b499-fd04eef80a15

23/05/25 12:27:09 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB

23/05/25 12:27:09 INFO spark.SparkEnv: Registering OutputCommitCoordinator

23/05/25 12:27:09 INFO spark.SparkContext: Added JAR [file:/home/raghavendr48edu/training\_project3/target/scala-2.11/sparkme-project\_2.11-1.0.jar](file:///\\home\raghavendr48edu\training_project3\target\scala-2.11\sparkme-project_2.11-1.0.jar) at spark://i

p-10-1-1-204.ap-south-1.compute.internal:40066/jars/sparkme-project\_2.11-1.0.jar with timestamp 1685017629956

23/05/25 12:27:10 INFO util.Utils: Using initial executors = 0, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spa

rk.executor.instances

23/05/25 12:27:10 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/25 12:27:10 INFO yarn.Client: Requesting a new application from cluster with 3 NodeManagers

23/05/25 12:27:10 INFO conf.Configuration: resource-types.xml not found

23/05/25 12:27:10 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/25 12:27:10 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (20480 MB per containe

r)

23/05/25 12:27:10 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead

23/05/25 12:27:10 INFO yarn.Client: Setting up container launch context for our AM

23/05/25 12:27:10 INFO yarn.Client: Setting up the launch environment for our AM container

23/05/25 12:27:10 INFO yarn.Client: Preparing resources for our AM container

23/05/25 12:27:11 INFO yarn.Client: Uploading resource [file:/tmp/spark-149fb59c-5b64-4b50-a117-6212d417e2c9/\_\_spark\_conf\_\_4441404531900358458.zip](file:///\\tmp\spark-149fb59c-5b64-4b50-a117-6212d417e2c9\__spark_conf__4441404531900358458.zip) -> hdfs://n

ameservice1/user/raghavendr48edu/.sparkStaging/application\_1684866872278\_1272/\_\_spark\_conf\_\_.zip

23/05/25 12:27:11 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/25 12:27:11 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/25 12:27:11 INFO spark.SecurityManager: Changing view acls groups to:

23/05/25 12:27:11 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/25 12:27:11 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu);

groups with view permissions: Set(); users with modify permissions: Set(raghavendr48edu); groups with modify permissions: Set()

23/05/25 12:27:11 INFO yarn.Client: Submitting application application\_1684866872278\_1272 to ResourceManager

23/05/25 12:27:11 INFO impl.YarnClientImpl: Submitted application application\_1684866872278\_1272

23/05/25 12:27:12 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:12 INFO yarn.Client:

client token: N/A

diagnostics: AM container is launched, waiting for AM container to Register with RM

ApplicationMaster host: N/A

ApplicationMaster RPC port: -1

queue: root.default

start time: 1685017631490

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_1272/>

user: raghavendr48edu

23/05/25 12:27:13 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:14 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:15 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:16 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:17 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:18 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:19 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: ACCEPTED)

23/05/25 12:27:20 INFO yarn.Client: Application report for application\_1684866872278\_1272 (state: RUNNING)

23/05/25 12:27:20 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: 10.1.2.40

ApplicationMaster RPC port: -1

queue: root.default

start time: 1685017631490

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_1272/>

user: raghavendr48edu

23/05/25 12:27:20 INFO cluster.YarnClientSchedulerBackend: Application application\_1684866872278\_1272 has started running.

23/05/25 12:27:20 INFO cluster.SchedulerExtensionServices: Starting Yarn extension services with app application\_1684866872278\_1272 and attemptId None

23/05/25 12:27:20 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 37637.

23/05/25 12:27:20 INFO netty.NettyBlockTransferService: Server created on ip-10-1-1-204.ap-south-1.compute.internal:37637

23/05/25 12:27:20 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy

23/05/25 12:27:20 INFO cluster.YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY\_HOSTS -> i

p-10-1-1-204.ap-south-1.compute.internal, PROXY\_URI\_BASES ->  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_1272>), /pr

oxy/application\_1684866872278\_1272

23/05/25 12:27:20 INFO storage.BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 37637, None)

23/05/25 12:27:20 INFO storage.BlockManagerMasterEndpoint: Registering block manager ip-10-1-1-204.ap-south-1.compute.internal:37637 with 366.3 MB RAM, Block

ManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 37637, None)

23/05/25 12:27:20 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 37637, None)

23/05/25 12:27:20 INFO storage.BlockManager: external shuffle service port = 7337

23/05/25 12:27:20 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 37637, None)

23/05/25 12:27:20 INFO util.log: Logging initialized @12715ms

23/05/25 12:27:21 INFO scheduler.EventLoggingListener: Logging events to hdfs://nameservice1/user/spark/applicationHistory/application\_1684866872278\_1272

23/05/25 12:27:21 INFO util.Utils: Using initial executors = 0, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spa

rk.executor.instances

23/05/25 12:27:21 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!

23/05/25 12:27:21 INFO spark.SparkContext: Registered listener com.cloudera.spark.lineage.NavigatorAppListener

23/05/25 12:27:21 INFO logging.DriverLogger$DfsAsyncWriter: Started driver log file sync to: /user/spark/driverLogs/application\_1684866872278\_1272\_driver.log

23/05/25 12:27:21 INFO cluster.YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.8

23/05/25 12:27:21 INFO internal.SharedState: loading hive config file: [file:/etc/hive/conf.cloudera.hive/hive-site.xml](file:///\\etc\hive\conf.cloudera.hive\hive-site.xml)

23/05/25 12:27:21 INFO internal.SharedState: spark.sql.warehouse.dir is not set, but hive.metastore.warehouse.dir is set. Setting spark.sql.warehouse.dir to

the value of hive.metastore.warehouse.dir ('/user/hive/warehouse').

23/05/25 12:27:21 INFO internal.SharedState: Warehouse path is '/user/hive/warehouse'.

23/05/25 12:27:21 INFO state.StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint

23/05/25 12:27:23 INFO datasources.FileSourceStrategy: Pruning directories with:

23/05/25 12:27:23 INFO datasources.FileSourceStrategy: Post-Scan Filters:

23/05/25 12:27:23 INFO datasources.FileSourceStrategy: Output Data Schema: struct<value: string>

23/05/25 12:27:23 INFO execution.FileSourceScanExec: Pushed Filters:

23/05/25 12:27:23 INFO cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)

23/05/25 12:27:24 INFO codegen.CodeGenerator: Code generated in 183.770987 ms

23/05/25 12:27:24 INFO memory.MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 336.4 KB, free 366.0 MB)

23/05/25 12:27:24 INFO memory.MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 32.2 KB, free 365.9 MB)

23/05/25 12:27:24 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 32.2 KB, free:

366.3 MB)

23/05/25 12:27:24 INFO spark.SparkContext: Created broadcast 0 from rdd at SparkSort.scala:23

23/05/25 12:27:24 INFO execution.FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 byt

es.

23/05/25 12:27:24 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:24 INFO scheduler.DAGScheduler: Got job 0 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:24 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (foreach at SparkSort.scala:33)

23/05/25 12:27:24 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:24 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:24 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:24 INFO memory.MemoryStore: Block broadcast\_1 stored as values in memory (estimated size 10.6 KB, free 365.9 MB)

23/05/25 12:27:24 INFO memory.MemoryStore: Block broadcast\_1\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.9 MB)

23/05/25 12:27:24 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.3 MB)

23/05/25 12:27:24 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:24 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:24 INFO cluster.YarnScheduler: Adding task set 0.0 with 1 tasks

23/05/25 12:27:24 INFO spark.ExecutorAllocationManager: Requesting 1 new executor because tasks are backlogged (new desired total will be 1)

23/05/25 12:27:35 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (10.1.2.40:54824) wi

th ID 1

23/05/25 12:27:35 INFO spark.ExecutorAllocationManager: New executor 1 has registered (new total is 1)

23/05/25 12:27:35 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, NO

DE\_LOCAL, 8347 bytes)

23/05/25 12:27:35 INFO storage.BlockManagerMasterEndpoint: Registering block manager ip-10-1-2-40.ap-south-1.compute.internal:33447 with 366.3 MB RAM, BlockM

anagerId(1, ip-10-1-2-40.ap-south-1.compute.internal, 33447, None)

23/05/25 12:27:36 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.3 MB)

23/05/25 12:27:46 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 32.2 KB, free: 3

66.3 MB)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added rdd\_4\_0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 32.0 B, free: 366.3 MB)

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 15041 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1)

(1/1)

23/05/25 12:27:50 INFO cluster.YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool

23/05/25 12:27:50 INFO scheduler.DAGScheduler: ResultStage 0 (foreach at SparkSort.scala:33) finished in 25.723 s

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Job 0 finished: foreach at SparkSort.scala:33, took 25.772812 s

-->Iteration 1 took 25786 ms

23/05/25 12:27:50 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Got job 1 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (foreach at SparkSort.scala:33)

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_2 stored as values in memory (estimated size 10.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_2\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.3 MB)

23/05/25 12:27:50 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:50 INFO cluster.YarnScheduler: Adding task set 1.0 with 1 tasks

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.3 MB)

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 244 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1

/1)

23/05/25 12:27:50 INFO cluster.YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool

23/05/25 12:27:50 INFO scheduler.DAGScheduler: ResultStage 1 (foreach at SparkSort.scala:33) finished in 0.276 s

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Job 1 finished: foreach at SparkSort.scala:33, took 0.282008 s

-->Iteration 2 took 288 ms

23/05/25 12:27:50 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Got job 2 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Final stage: ResultStage 2 (foreach at SparkSort.scala:33)

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_3 stored as values in memory (estimated size 10.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_3\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.3 MB)

23/05/25 12:27:50 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:50 INFO cluster.YarnScheduler: Adding task set 2.0 with 1 tasks

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.3 MB)

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 95 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1/

1)

23/05/25 12:27:50 INFO cluster.YarnScheduler: Removed TaskSet 2.0, whose tasks have all completed, from pool

23/05/25 12:27:50 INFO scheduler.DAGScheduler: ResultStage 2 (foreach at SparkSort.scala:33) finished in 0.111 s

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Job 2 finished: foreach at SparkSort.scala:33, took 0.118225 s

-->Iteration 3 took 125 ms

23/05/25 12:27:50 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Got job 3 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Final stage: ResultStage 3 (foreach at SparkSort.scala:33)

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_4 stored as values in memory (estimated size 10.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_4\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.2 MB)

23/05/25 12:27:50 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:50 INFO cluster.YarnScheduler: Adding task set 3.0 with 1 tasks

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.2 MB)

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 70 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1/

1)

23/05/25 12:27:50 INFO cluster.YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool

23/05/25 12:27:50 INFO scheduler.DAGScheduler: ResultStage 3 (foreach at SparkSort.scala:33) finished in 0.083 s

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Job 3 finished: foreach at SparkSort.scala:33, took 0.088030 s

-->Iteration 4 took 93 ms

23/05/25 12:27:50 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Got job 4 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Final stage: ResultStage 4 (foreach at SparkSort.scala:33)

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting ResultStage 4 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_5 stored as values in memory (estimated size 10.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_5\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_5\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.2 MB)

23/05/25 12:27:50 INFO spark.SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 4 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:50 INFO cluster.YarnScheduler: Adding task set 4.0 with 1 tasks

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 4.0 (TID 4, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_5\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.2 MB)

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 4.0 (TID 4) in 104 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1

/1)

23/05/25 12:27:50 INFO cluster.YarnScheduler: Removed TaskSet 4.0, whose tasks have all completed, from pool

23/05/25 12:27:50 INFO scheduler.DAGScheduler: ResultStage 4 (foreach at SparkSort.scala:33) finished in 0.115 s

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Job 4 finished: foreach at SparkSort.scala:33, took 0.117970 s

-->Iteration 5 took 123 ms

23/05/25 12:27:50 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Got job 5 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Final stage: ResultStage 5 (foreach at SparkSort.scala:33)

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_6 stored as values in memory (estimated size 10.6 KB, free 365.9 MB)

23/05/25 12:27:50 INFO memory.MemoryStore: Block broadcast\_6\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.8 MB)

23/05/25 12:27:50 INFO storage.BlockManagerInfo: Added broadcast\_6\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.2 MB)

23/05/25 12:27:50 INFO spark.SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:50 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:50 INFO cluster.YarnScheduler: Adding task set 5.0 with 1 tasks

23/05/25 12:27:50 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 5.0 (TID 5, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_6\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.2 MB)

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 5.0 (TID 5) in 493 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1

/1)

23/05/25 12:27:51 INFO cluster.YarnScheduler: Removed TaskSet 5.0, whose tasks have all completed, from pool

23/05/25 12:27:51 INFO scheduler.DAGScheduler: ResultStage 5 (foreach at SparkSort.scala:33) finished in 0.506 s

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Job 5 finished: foreach at SparkSort.scala:33, took 0.510316 s

-->Iteration 6 took 516 ms

23/05/25 12:27:51 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Got job 6 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Final stage: ResultStage 6 (foreach at SparkSort.scala:33)

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting ResultStage 6 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_7 stored as values in memory (estimated size 10.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_7\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_7\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.2 MB)

23/05/25 12:27:51 INFO spark.SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 6 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:51 INFO cluster.YarnScheduler: Adding task set 6.0 with 1 tasks

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 6.0 (TID 6, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_7\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.2 MB)

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 6.0 (TID 6) in 94 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1/

1)

23/05/25 12:27:51 INFO cluster.YarnScheduler: Removed TaskSet 6.0, whose tasks have all completed, from pool

23/05/25 12:27:51 INFO scheduler.DAGScheduler: ResultStage 6 (foreach at SparkSort.scala:33) finished in 0.108 s

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Job 6 finished: foreach at SparkSort.scala:33, took 0.110870 s

-->Iteration 7 took 115 ms

23/05/25 12:27:51 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Got job 7 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Final stage: ResultStage 7 (foreach at SparkSort.scala:33)

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting ResultStage 7 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_8 stored as values in memory (estimated size 10.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_8\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_8\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.2 MB)

23/05/25 12:27:51 INFO spark.SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 7 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:51 INFO cluster.YarnScheduler: Adding task set 7.0 with 1 tasks

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 7.0 (TID 7, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_8\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.2 MB)

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 7.0 (TID 7) in 70 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1/

1)

23/05/25 12:27:51 INFO cluster.YarnScheduler: Removed TaskSet 7.0, whose tasks have all completed, from pool

23/05/25 12:27:51 INFO scheduler.DAGScheduler: ResultStage 7 (foreach at SparkSort.scala:33) finished in 0.081 s

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Job 7 finished: foreach at SparkSort.scala:33, took 0.084157 s

-->Iteration 8 took 89 ms

23/05/25 12:27:51 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Got job 8 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Final stage: ResultStage 8 (foreach at SparkSort.scala:33)

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting ResultStage 8 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_9 stored as values in memory (estimated size 10.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_9\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_9\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free: 3

66.2 MB)

23/05/25 12:27:51 INFO spark.SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 8 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:51 INFO cluster.YarnScheduler: Adding task set 8.0 with 1 tasks

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 8.0 (TID 8, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_9\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 36

6.2 MB)

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 8.0 (TID 8) in 60 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1/

1)

23/05/25 12:27:51 INFO cluster.YarnScheduler: Removed TaskSet 8.0, whose tasks have all completed, from pool

23/05/25 12:27:51 INFO scheduler.DAGScheduler: ResultStage 8 (foreach at SparkSort.scala:33) finished in 0.072 s

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Job 8 finished: foreach at SparkSort.scala:33, took 0.075601 s

-->Iteration 9 took 80 ms

23/05/25 12:27:51 INFO spark.SparkContext: Starting job: foreach at SparkSort.scala:33

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Got job 9 (foreach at SparkSort.scala:33) with 1 output partitions

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Final stage: ResultStage 9 (foreach at SparkSort.scala:33)

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting ResultStage 9 (MapPartitionsRDD[4] at map at SparkSort.scala:25), which has no missing parents

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_10 stored as values in memory (estimated size 10.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO memory.MemoryStore: Block broadcast\_10\_piece0 stored as bytes in memory (estimated size 5.6 KB, free 365.8 MB)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_10\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:37637 (size: 5.6 KB, free:

366.2 MB)

23/05/25 12:27:51 INFO spark.SparkContext: Created broadcast 10 from broadcast at DAGScheduler.scala:1164

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 9 (MapPartitionsRDD[4] at map at SparkSort.scala:25) (first 15 tas

ks are for partitions Vector(0))

23/05/25 12:27:51 INFO cluster.YarnScheduler: Adding task set 9.0 with 1 tasks

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 9.0 (TID 9, ip-10-1-2-40.ap-south-1.compute.internal, executor 1, partition 0, PR

OCESS\_LOCAL, 8347 bytes)

23/05/25 12:27:51 INFO storage.BlockManagerInfo: Added broadcast\_10\_piece0 in memory on ip-10-1-2-40.ap-south-1.compute.internal:33447 (size: 5.6 KB, free: 3

66.2 MB)

23/05/25 12:27:51 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 9.0 (TID 9) in 73 ms on ip-10-1-2-40.ap-south-1.compute.internal (executor 1) (1/

1)

23/05/25 12:27:51 INFO cluster.YarnScheduler: Removed TaskSet 9.0, whose tasks have all completed, from pool

23/05/25 12:27:51 INFO scheduler.DAGScheduler: ResultStage 9 (foreach at SparkSort.scala:33) finished in 0.085 s

23/05/25 12:27:51 INFO scheduler.DAGScheduler: Job 9 finished: foreach at SparkSort.scala:33, took 0.088958 s

-->Iteration 10 took 94 ms

23/05/25 12:27:51 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread

23/05/25 12:27:51 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors

23/05/25 12:27:51 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down

23/05/25 12:27:51 INFO cluster.SchedulerExtensionServices: Stopping SchedulerExtensionServices

(serviceOption=None,

services=List(),

started=false)

23/05/25 12:27:51 INFO cluster.YarnClientSchedulerBackend: Stopped

23/05/25 12:27:51 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!

23/05/25 12:27:51 INFO memory.MemoryStore: MemoryStore cleared

23/05/25 12:27:51 INFO storage.BlockManager: BlockManager stopped

23/05/25 12:27:51 INFO storage.BlockManagerMaster: BlockManagerMaster stopped

23/05/25 12:27:51 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

23/05/25 12:27:51 INFO spark.SparkContext: Successfully stopped SparkContext

23/05/25 12:27:51 INFO util.ShutdownHookManager: Shutdown hook called

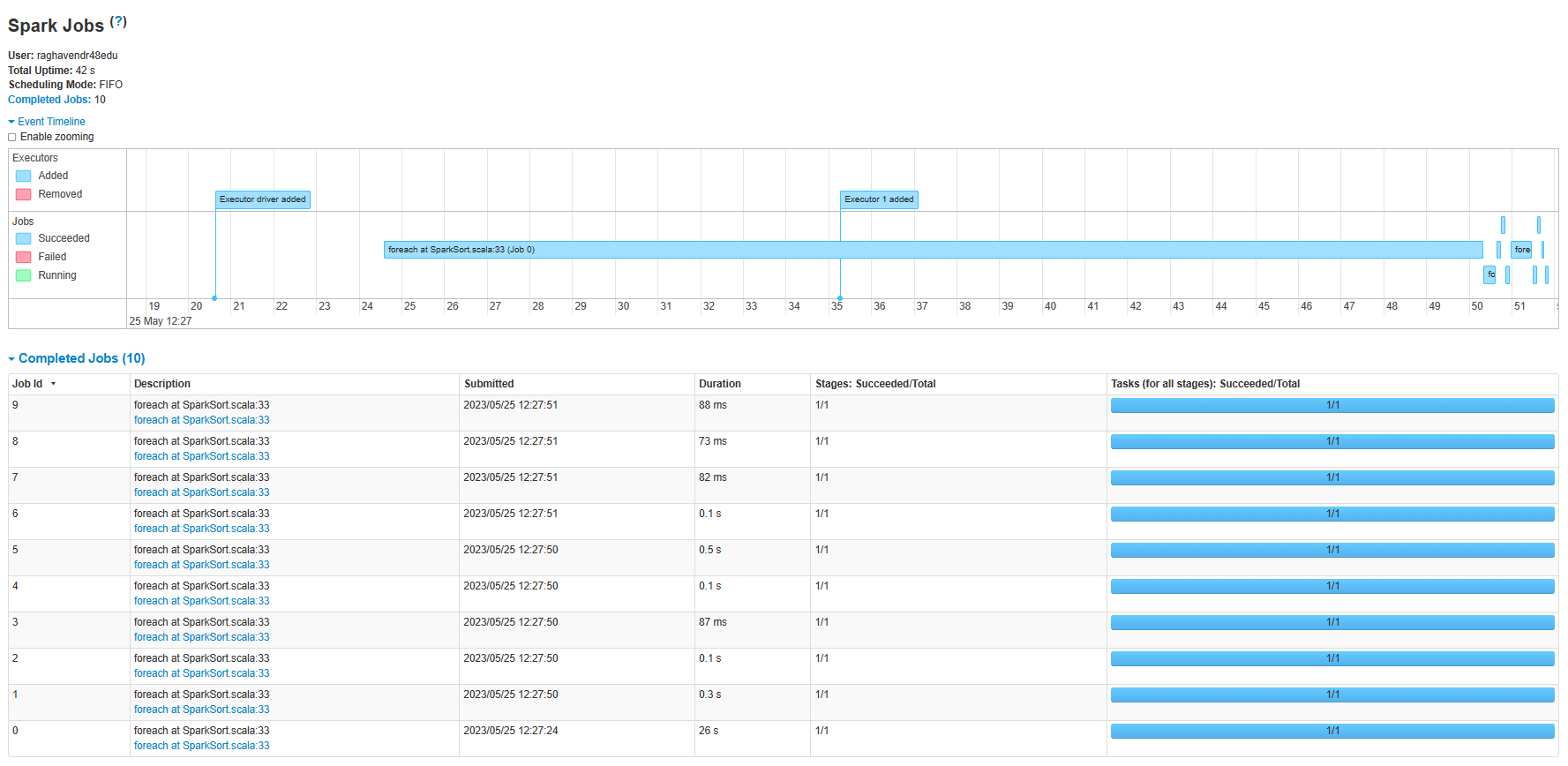
23/05/25 12:27:51 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-149fb59c-5b64-4b50-a117-6212d417e2c9

23/05/25 12:27:51 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-aeca6266-9a43-4039-b421-4d6e67bad587

[raghavendr48edu@ip-10-1-1-204 training\_project3]$

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

4.Analyze the behavior of spark application on Spark web UI



OR

**HDFSTest.scala**

package org.apache.spark.examples

import org.apache.spark.sql.SparkSession

object HdfsTest {

/\*\* Usage: HdfsTest [file] \*/  
 def main(args: Array[String]) {  
 if (args.length < 1) {  
 System.err.println("Usage: HdfsTest <file>")  
 System.exit(1)  
 }  
 val spark = SparkSession  
 .builder  
 .appName("HdfsTest")  
 .getOrCreate()  
 val file = spark.read.text(args(0)).rdd  
 val mapped = file.map(s => s.length).cache()  
 for (iter <- 1 to 10) {  
 val start = System.currentTimeMillis()  
 for (x <- mapped) { x + 2 }  
 val end = System.currentTimeMillis()  
 println("Iteration " + iter + " took " + (end-start) + " ms")  
 }  
 spark.stop()  
 }  
}

[raghavendr48edu@ip-10-1-1-204 ~]$ **spark-submit --class org.apache.spark.examples.HdfsTest --master yarn --deploy-mode cluster /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/examples/jar**

**s/spark-examples\*.jar use\_cases/HdfsTest.scala**

23/05/23 14:49:42 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/23 14:49:42 INFO yarn.Client: Requesting a new application from cluster with 3 NodeManagers

23/05/23 14:49:42 INFO conf.Configuration: resource-types.xml not found

23/05/23 14:49:42 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/23 14:49:42 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (20480 MB per container)

23/05/23 14:49:42 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead

23/05/23 14:49:42 INFO yarn.Client: Setting up container launch context for our AM

23/05/23 14:49:42 INFO yarn.Client: Setting up the launch environment for our AM container

23/05/23 14:49:42 INFO yarn.Client: Preparing resources for our AM container

23/05/23 14:49:42 INFO yarn.Client: Uploading resource [file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/examples/jars/spark-examples\_2.11-2.4.0-cdh6.2.1.jar](file:///\\opt\cloudera\parcels\CDH-6.2.1-1.cdh6.2.1.p0.1425774\lib\spark\examples\jars\spark-examples_2.11-2.4.0-cdh6.2.1.jar) -> hdfs://nameservice1/user/ragh

avendr48edu/.sparkStaging/application\_1684298513961\_1256/spark-examples\_2.11-2.4.0-cdh6.2.1.jar

23/05/23 14:49:43 INFO yarn.Client: Uploading resource [file:/tmp/spark-8e35dab9-211c-491e-8347-b29792b53503/\_\_spark\_conf\_\_8123679462684054128.zip](file:///\\tmp\spark-8e35dab9-211c-491e-8347-b29792b53503\__spark_conf__8123679462684054128.zip) -> hdfs://nameservice1/user/raghavendr48edu/.sparkStaging/appli

cation\_1684298513961\_1256/\_\_spark\_conf\_\_.zip

23/05/23 14:49:43 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/23 14:49:43 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/23 14:49:43 INFO spark.SecurityManager: Changing view acls groups to:

23/05/23 14:49:43 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/23 14:49:43 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu); groups with view permissions: Set(); users with mo

dify permissions: Set(raghavendr48edu); groups with modify permissions: Set()

23/05/23 14:49:43 INFO conf.HiveConf: Found configuration file [file:/etc/hive/conf.cloudera.hive/hive-site.xml](file:///\\etc\hive\conf.cloudera.hive\hive-site.xml)

23/05/23 14:49:43 INFO security.YARNHadoopDelegationTokenManager: Attempting to load user's ticket cache.

23/05/23 14:49:43 INFO yarn.Client: Submitting application application\_1684298513961\_1256 to ResourceManager

23/05/23 14:49:43 INFO impl.YarnClientImpl: Submitted application application\_1684298513961\_1256

23/05/23 14:49:44 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:44 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: N/A

ApplicationMaster RPC port: -1

queue: root.default

start time: 1684853383303

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_1256/>

user: raghavendr48edu

23/05/23 14:49:45 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:46 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:47 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:48 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:49 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:50 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:51 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:52 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:53 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:54 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:55 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:56 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:57 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:58 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:49:59 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: ACCEPTED)

23/05/23 14:50:00 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:00 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: ip-10-1-2-40.ap-south-1.compute.internal

ApplicationMaster RPC port: 36414

queue: root.default

start time: 1684853383303

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_1256/>

user: raghavendr48edu

23/05/23 14:50:01 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:02 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:03 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:04 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:05 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:06 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:07 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:08 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:09 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:10 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:11 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:12 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:13 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:14 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:15 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:16 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:17 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:18 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:19 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:20 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:21 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:22 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:23 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:24 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:25 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:26 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: RUNNING)

23/05/23 14:50:27 INFO yarn.Client: Application report for application\_1684298513961\_1256 (state: FINISHED)

23/05/23 14:50:27 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: ip-10-1-2-40.ap-south-1.compute.internal

ApplicationMaster RPC port: 36414

queue: root.default

start time: 1684853383303

final status: SUCCEEDED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_1256/>

user: raghavendr48edu

23/05/23 14:50:27 INFO util.ShutdownHookManager: Shutdown hook called

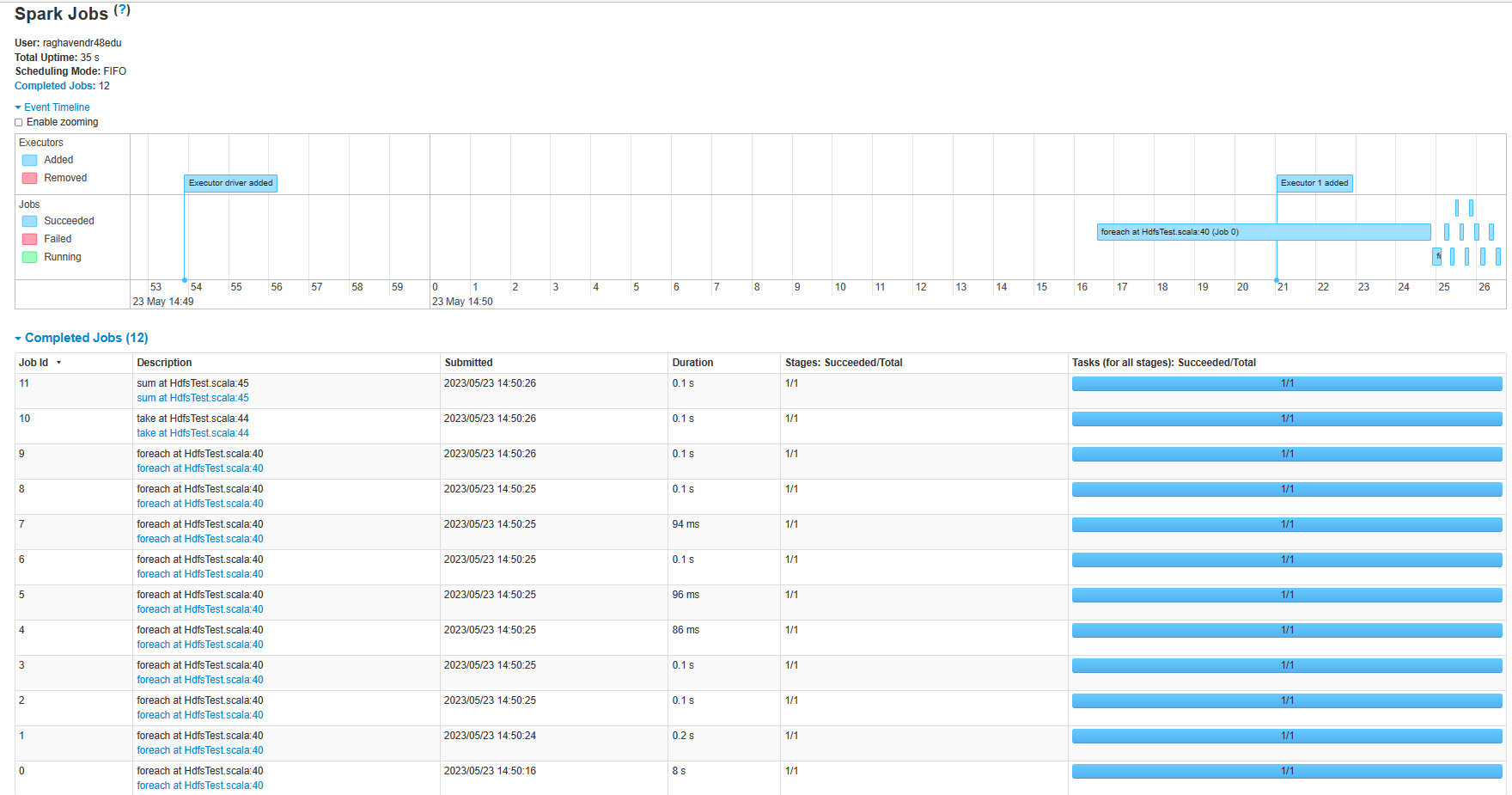
23/05/23 14:50:27 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-cf98be70-646d-495c-9c18-bc8c8565c63f

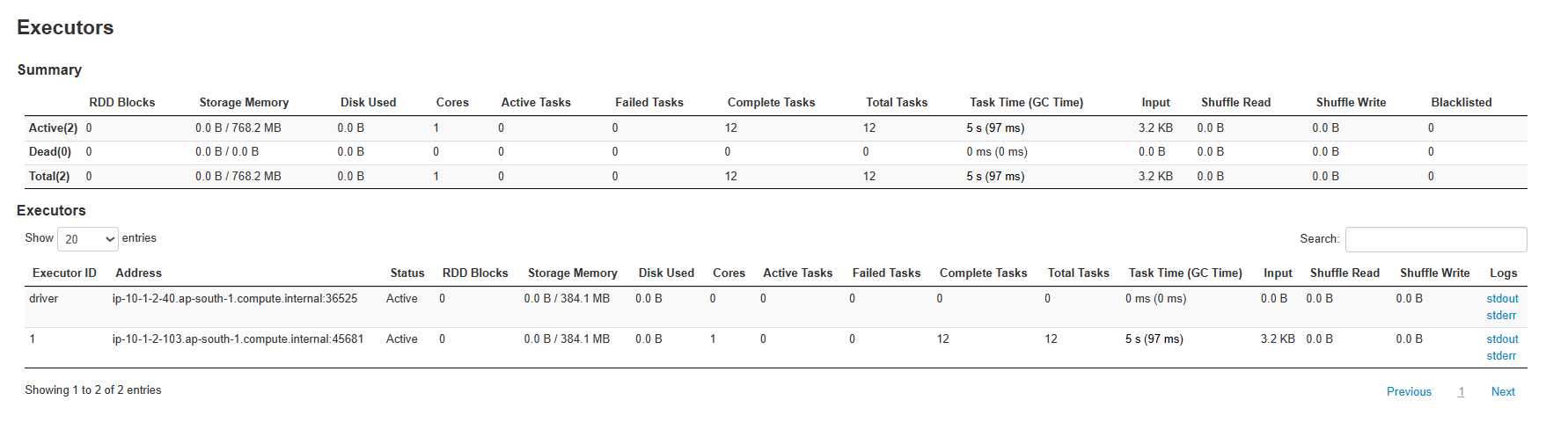
23/05/23 14:50:27 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-8e35dab9-211c-491e-8347-b29792b53503

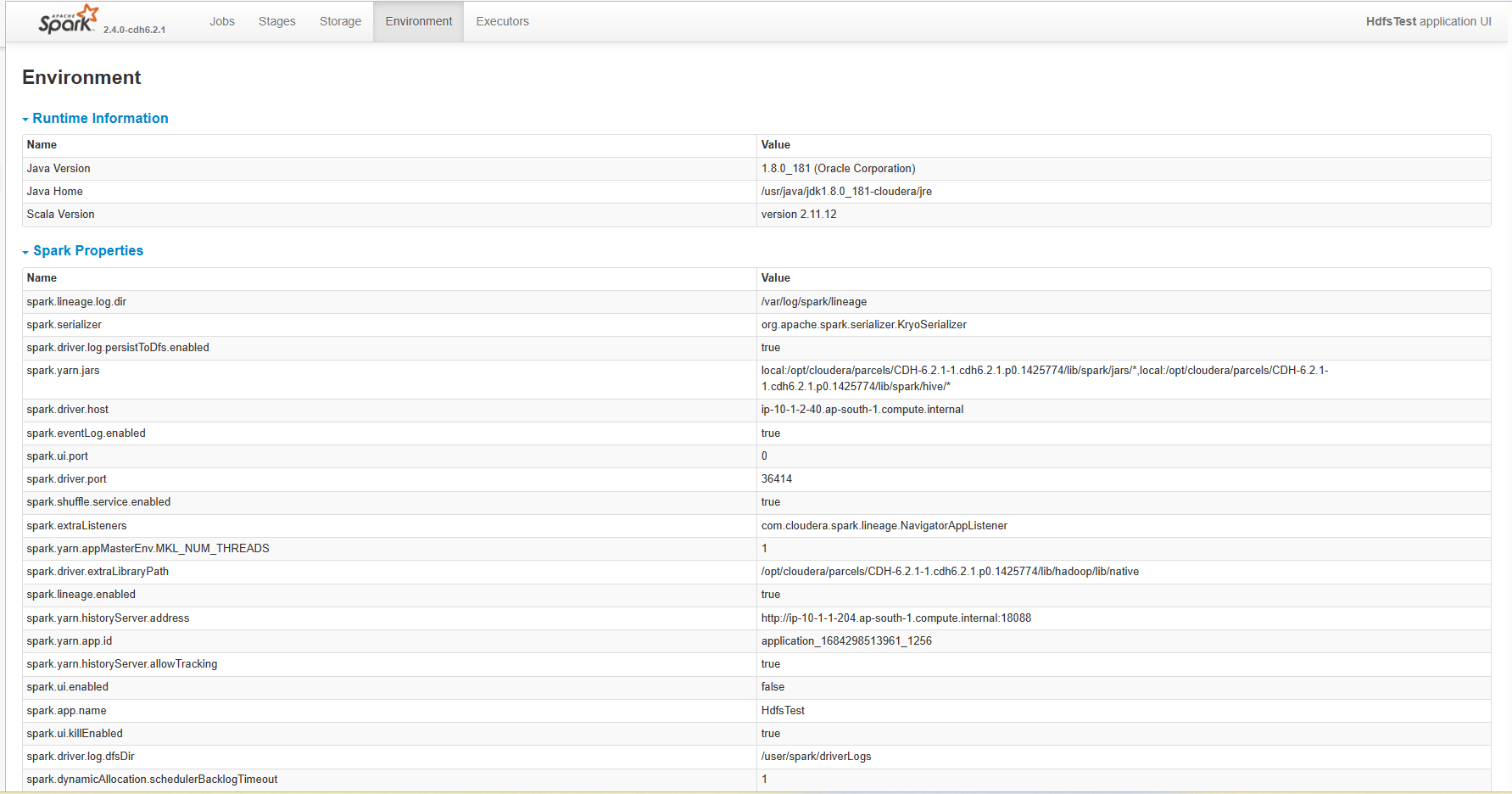
[raghavendr48edu@ip-10-1-1-204 ~]$

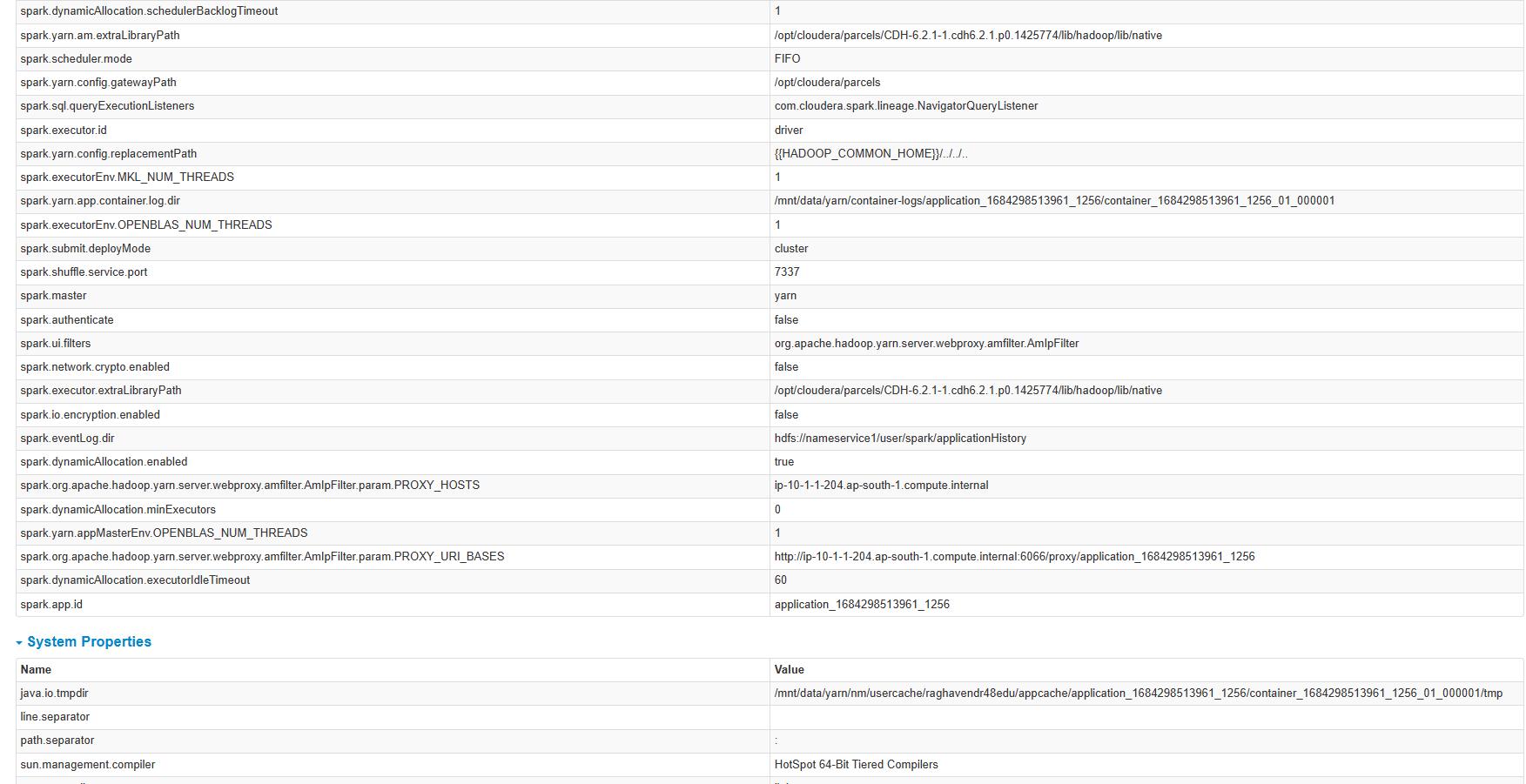
4.Analyze the behavior of spark application on Spark web UI

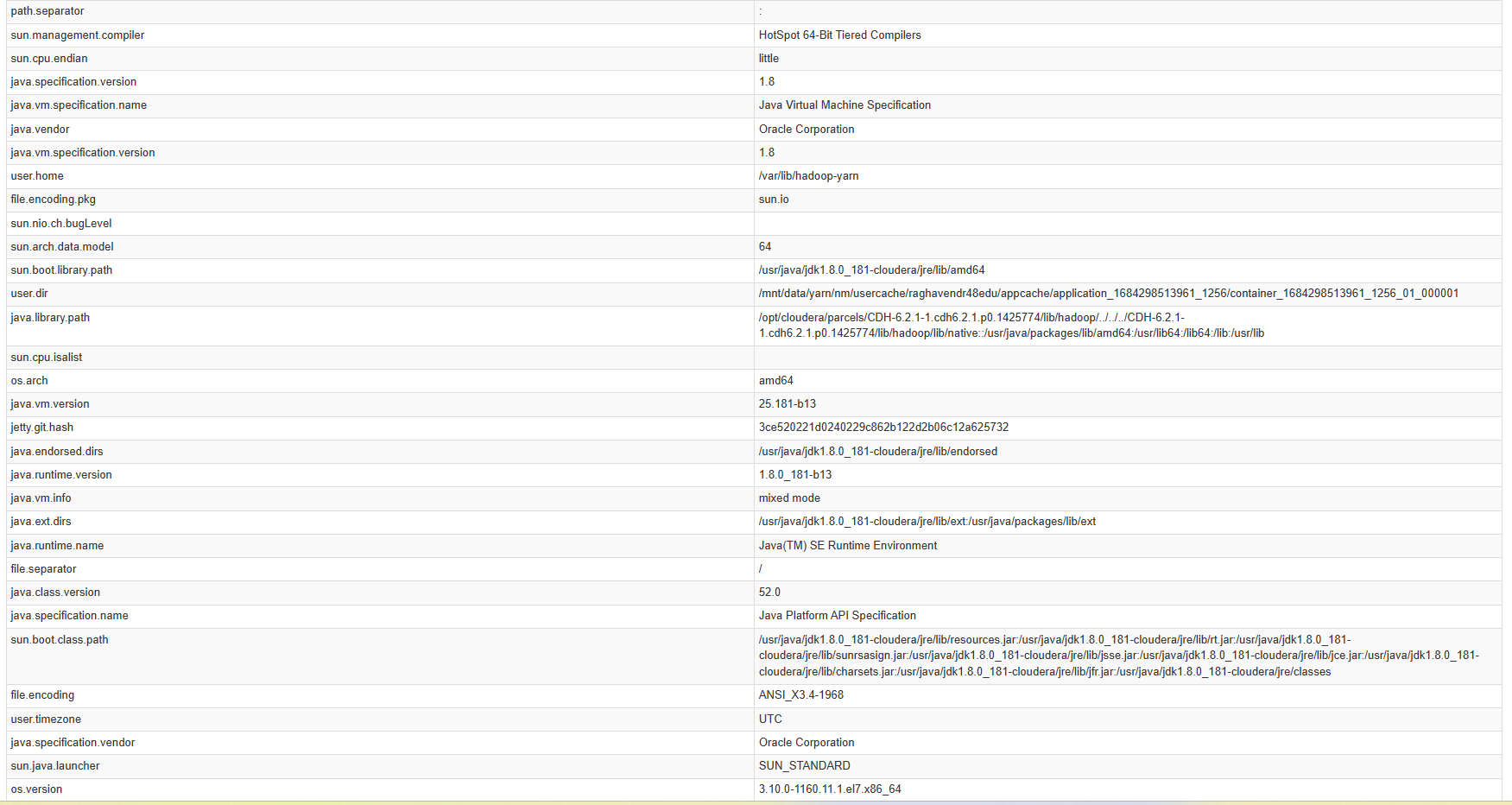
*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

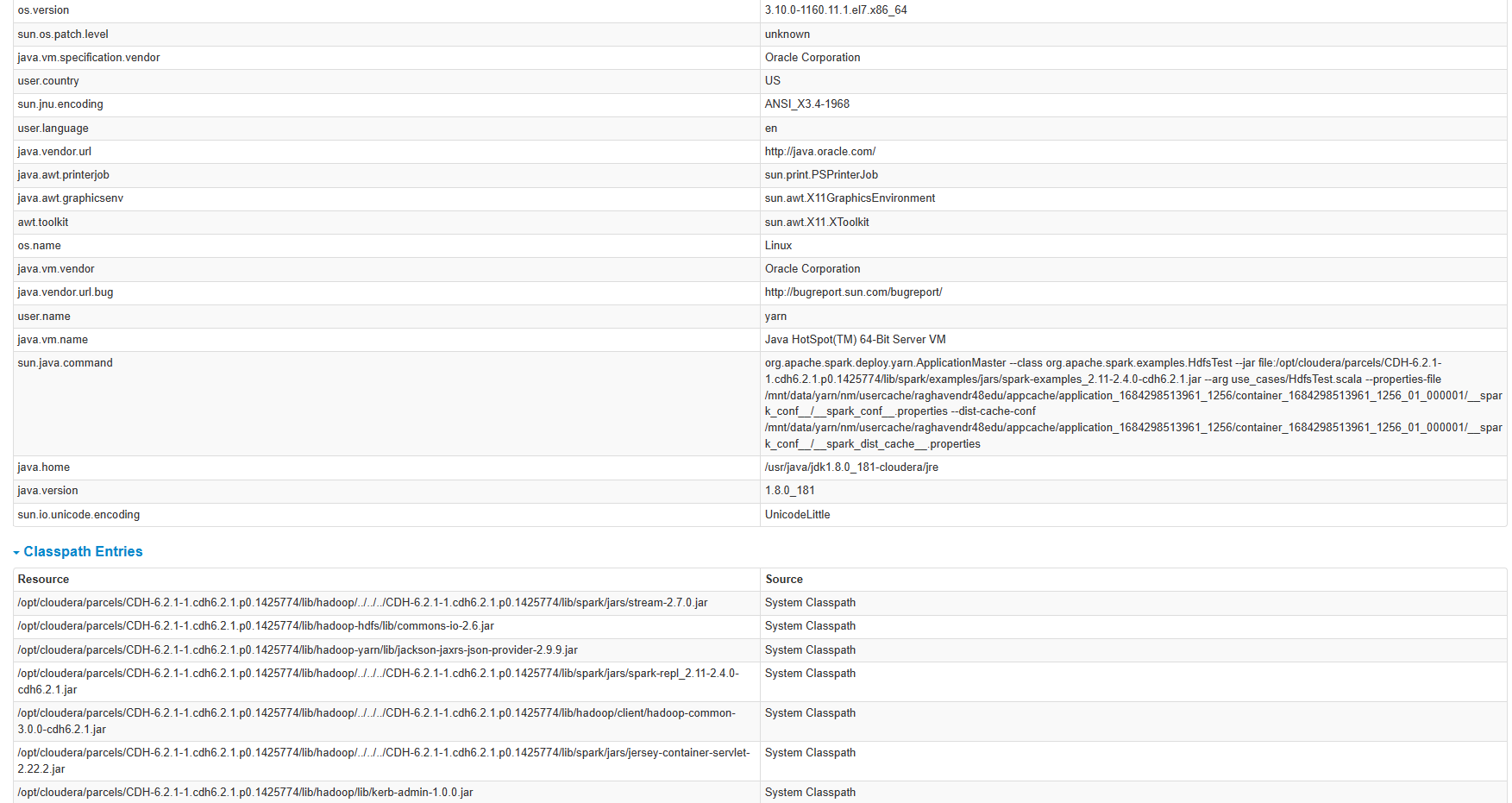












5.Add custom logs in your application and re-execute the application. Once

executed check the Spark logs.

a. Modify the script ‘sort.py’ to add a custom logger (first screenshot)

log4jLogger = spark.\_jvm.org.apache.log4j

LOGGER = log4jLogger.LogManager.getLogger(\_\_name\_\_)   
LOGGER.info(“pyspark custom logger added by Nandini”)

b. Re-submit the spark job with updated script

c. View the yarn logs yarn logs --applicationId

from \_\_future\_\_ import print\_function

import sys,time

from pyspark.sql import SparkSession

if \_\_name\_\_ == "\_\_main\_\_":

if len(sys.argv) != 2:

print("Usage: hdfstest.py <file>", file=sys.stderr)

exit(-1)

spark = SparkSession\

.builder\

.appName("HdfsTest\_py")\

.getOrCreate()

file\_ = spark.read.text(sys.argv[1]).rdd

**log4jLogger = spark.\_jvm.org.apache.log4j**

**LOGGER = log4jLogger.LogManager.getLogger(\_\_name\_\_)**

**LOGGER.info(“pyspark custom logger added by Raghav”)**

mapped = file\_.map(lambda s:len(s)).cache()

for i in range(10):

start\_time = time.time()

mapped.map(lambda x: x+2)

end\_time = time.time()

print('----> Iteration took:', end\_time - start\_time,'ms')

**result=mapped.take(10)**

print(result)

spark.stop()

[raghavendr48edu@ip-10-1-1-204 ~]$ spark-submit sort.py input\_sort\_py.txt

23/05/26 07:19:12 INFO spark.SparkContext: Running Spark version 2.4.0-cdh6.2.1

23/05/26 07:19:12 INFO logging.DriverLogger: Added a local log appender at: /tmp/spark-6cdab5ed-bbba-4c2c-a7ec-441a4b64277d/\_\_driver\_logs\_\_/driver.log

23/05/26 07:19:12 INFO spark.SparkContext: Submitted application: HdfsTest\_py

23/05/26 07:19:12 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/26 07:19:12 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/26 07:19:12 INFO spark.SecurityManager: Changing view acls groups to:

23/05/26 07:19:12 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/26 07:19:12 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu); groups with view permissions: Set(); users with modify permissions: Set(ragh

avendr48edu); groups with modify permissions: Set()

23/05/26 07:19:12 INFO util.Utils: Successfully started service 'sparkDriver' on port 41949.

23/05/26 07:19:12 INFO spark.SparkEnv: Registering MapOutputTracker

23/05/26 07:19:12 INFO spark.SparkEnv: Registering BlockManagerMaster

23/05/26 07:19:12 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information

23/05/26 07:19:12 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up

23/05/26 07:19:12 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-8df5ae21-54a7-4251-9ed4-77299c9f2caa

23/05/26 07:19:12 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB

23/05/26 07:19:12 INFO spark.SparkEnv: Registering OutputCommitCoordinator

23/05/26 07:19:12 INFO util.Utils: Using initial executors = 0, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances

23/05/26 07:19:13 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/26 07:19:13 INFO yarn.Client: Requesting a new application from cluster with 3 NodeManagers

23/05/26 07:19:13 INFO conf.Configuration: resource-types.xml not found

23/05/26 07:19:13 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/26 07:19:13 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (20480 MB per container)

23/05/26 07:19:13 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead

23/05/26 07:19:13 INFO yarn.Client: Setting up container launch context for our AM

23/05/26 07:19:13 INFO yarn.Client: Setting up the launch environment for our AM container

23/05/26 07:19:13 INFO yarn.Client: Preparing resources for our AM container

23/05/26 07:19:13 INFO yarn.Client: Uploading resource [file:/tmp/spark-6cdab5ed-bbba-4c2c-a7ec-441a4b64277d/\_\_spark\_conf\_\_4653700732432819453.zip](file:///\\tmp\spark-6cdab5ed-bbba-4c2c-a7ec-441a4b64277d\__spark_conf__4653700732432819453.zip) -> hdfs://nameservice1/user/raghavendr48edu/.sparkStaging/application\_1684866872278\_2046/

\_\_spark\_conf\_\_.zip

23/05/26 07:19:14 INFO spark.SecurityManager: Changing view acls to: raghavendr48edu

23/05/26 07:19:14 INFO spark.SecurityManager: Changing modify acls to: raghavendr48edu

23/05/26 07:19:14 INFO spark.SecurityManager: Changing view acls groups to:

23/05/26 07:19:14 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/26 07:19:14 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(raghavendr48edu); groups with view permissions: Set(); users with modify permissions: Set(ragh

avendr48edu); groups with modify permissions: Set()

23/05/26 07:19:14 INFO yarn.Client: Submitting application application\_1684866872278\_2046 to ResourceManager

23/05/26 07:19:14 INFO impl.YarnClientImpl: Submitted application application\_1684866872278\_2046

23/05/26 07:19:15 INFO yarn.Client: Application report for application\_1684866872278\_2046 (state: ACCEPTED)

23/05/26 07:19:15 INFO yarn.Client:

client token: N/A

diagnostics: AM container is launched, waiting for AM container to Register with RM

ApplicationMaster host: N/A

ApplicationMaster RPC port: -1

queue: root.default

start time: 1685085554231

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_2046/>

user: raghavendr48edu

23/05/26 07:19:16 INFO yarn.Client: Application report for application\_1684866872278\_2046 (state: ACCEPTED)

23/05/26 07:19:17 INFO cluster.YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY\_HOSTS -> ip-10-1-1-204.ap-south-1.compute.internal, PROXY\_URI\_BASES -> [http://ip-10-1-1-](http://ip-10-1-1-/)

204.ap-south-1.compute.internal:6066/proxy/application\_1684866872278\_2046), /proxy/application\_1684866872278\_2046

23/05/26 07:19:17 INFO yarn.Client: Application report for application\_1684866872278\_2046 (state: RUNNING)

23/05/26 07:19:17 INFO yarn.Client:

client token: N/A

diagnostics: N/A

ApplicationMaster host: 10.1.2.103

ApplicationMaster RPC port: -1

queue: root.default

start time: 1685085554231

final status: UNDEFINED

         tracking URL:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_2046/>

user: raghavendr48edu

23/05/26 07:19:17 INFO cluster.YarnClientSchedulerBackend: Application application\_1684866872278\_2046 has started running.

23/05/26 07:19:17 INFO cluster.SchedulerExtensionServices: Starting Yarn extension services with app application\_1684866872278\_2046 and attemptId None

23/05/26 07:19:17 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 45048.

23/05/26 07:19:17 INFO netty.NettyBlockTransferService: Server created on ip-10-1-1-204.ap-south-1.compute.internal:45048

23/05/26 07:19:17 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy

23/05/26 07:19:17 INFO storage.BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 45048, None)

23/05/26 07:19:17 INFO storage.BlockManagerMasterEndpoint: Registering block manager ip-10-1-1-204.ap-south-1.compute.internal:45048 with 366.3 MB RAM, BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 45048, None)

23/05/26 07:19:17 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 45048, None)

23/05/26 07:19:17 INFO storage.BlockManager: external shuffle service port = 7337

23/05/26 07:19:17 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-10-1-1-204.ap-south-1.compute.internal, 45048, None)

23/05/26 07:19:17 INFO util.log: Logging initialized @7084ms

23/05/26 07:19:17 INFO scheduler.EventLoggingListener: Logging events to hdfs://nameservice1/user/spark/applicationHistory/application\_1684866872278\_2046

23/05/26 07:19:17 INFO util.Utils: Using initial executors = 0, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances

23/05/26 07:19:17 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!

23/05/26 07:19:17 INFO spark.SparkContext: Registered listener com.cloudera.spark.lineage.NavigatorAppListener

23/05/26 07:19:17 INFO logging.DriverLogger$DfsAsyncWriter: Started driver log file sync to: /user/spark/driverLogs/application\_1684866872278\_2046\_driver.log

23/05/26 07:19:17 INFO cluster.YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.8

23/05/26 07:19:17 INFO internal.SharedState: loading hive config file: [file:/etc/hive/conf.cloudera.hive/hive-site.xml](file:///\\etc\hive\conf.cloudera.hive\hive-site.xml)

23/05/26 07:19:18 INFO internal.SharedState: spark.sql.warehouse.dir is not set, but hive.metastore.warehouse.dir is set. Setting spark.sql.warehouse.dir to the value of hive.metastore.warehouse.dir ('/user/hive/warehouse').

23/05/26 07:19:18 INFO internal.SharedState: Warehouse path is '/user/hive/warehouse'.

23/05/26 07:19:18 INFO cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)

23/05/26 07:19:18 INFO state.StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint

23/05/26 07:19:20 INFO datasources.FileSourceStrategy: Pruning directories with:

23/05/26 07:19:20 INFO datasources.FileSourceStrategy: Post-Scan Filters:

23/05/26 07:19:20 INFO datasources.FileSourceStrategy: Output Data Schema: struct<value: string>

23/05/26 07:19:20 INFO execution.FileSourceScanExec: Pushed Filters:

23/05/26 07:19:20 INFO codegen.CodeGenerator: Code generated in 179.628026 ms

23/05/26 07:19:20 INFO memory.MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 337.3 KB, free 366.0 MB)

23/05/26 07:19:20 INFO memory.MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 32.3 KB, free 365.9 MB)

23/05/26 07:19:20 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 32.3 KB, free: 366.3 MB)

23/05/26 07:19:21 INFO spark.SparkContext: Created broadcast 0 from javaToPython at NativeMethodAccessorImpl.java:0

23/05/26 07:19:21 INFO execution.FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes.

23/05/26 07:19:21 INFO \_\_main\_\_: **pyspark custom logger added by Raghav**

----> Iteration took: 2.193450927734375e-05 ms

23/05/26 07:19:21 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:21 INFO scheduler.DAGScheduler: Got job 0 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:21 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (runJob at PythonRDD.scala:153)

23/05/26 07:19:21 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:21 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:21 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (PythonRDD[5] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:21 INFO memory.MemoryStore: Block broadcast\_1 stored as values in memory (estimated size 11.4 KB, free 365.9 MB)

23/05/26 07:19:21 INFO memory.MemoryStore: Block broadcast\_1\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.9 MB)

23/05/26 07:19:21 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.3 MB)

23/05/26 07:19:21 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:21 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (PythonRDD[5] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:21 INFO cluster.YarnScheduler: Adding task set 0.0 with 1 tasks

23/05/26 07:19:21 INFO spark.ExecutorAllocationManager: Requesting 1 new executor because tasks are backlogged (new desired total will be 1)

23/05/26 07:19:24 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (10.1.2.103:38166) with ID 1

23/05/26 07:19:24 INFO spark.ExecutorAllocationManager: New executor 1 has registered (new total is 1)

23/05/26 07:19:24 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, NODE\_LOCAL, 8347 bytes)

23/05/26 07:19:24 INFO storage.BlockManagerMasterEndpoint: Registering block manager ip-10-1-2-103.ap-south-1.compute.internal:34192 with 366.3 MB RAM, BlockManagerId(1, ip-10-1-2-103.ap-south-1.compute.internal, 34192, None)

23/05/26 07:19:24 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.3 MB)

23/05/26 07:19:26 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 32.3 KB, free: 366.3 MB)

23/05/26 07:19:27 INFO storage.BlockManagerInfo: Added rdd\_4\_0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 69.0 B, free: 366.3 MB)

23/05/26 07:19:27 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 3341 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:27 INFO cluster.YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool

23/05/26 07:19:27 INFO python.PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 47806

23/05/26 07:19:27 INFO scheduler.DAGScheduler: ResultStage 0 (runJob at PythonRDD.scala:153) finished in 6.359 s

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:153, took 6.407377 s

[1, 1, 1]

----> Iteration took: 2.1696090698242188e-05 ms

23/05/26 07:19:27 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Got job 1 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (runJob at PythonRDD.scala:153)

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_2 stored as values in memory (estimated size 11.4 KB, free 365.9 MB)

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_2\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.9 MB)

23/05/26 07:19:27 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.3 MB)

23/05/26 07:19:27 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[6] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:27 INFO cluster.YarnScheduler: Adding task set 1.0 with 1 tasks

23/05/26 07:19:27 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:27 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.3 MB)

23/05/26 07:19:27 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 119 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:27 INFO cluster.YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool

23/05/26 07:19:27 INFO scheduler.DAGScheduler: ResultStage 1 (runJob at PythonRDD.scala:153) finished in 0.138 s

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Job 1 finished: runJob at PythonRDD.scala:153, took 0.145067 s

[1, 1, 1]

----> Iteration took: 2.0265579223632812e-05 ms

23/05/26 07:19:27 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Got job 2 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Final stage: ResultStage 2 (runJob at PythonRDD.scala:153)

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Submitting ResultStage 2 (PythonRDD[7] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_3 stored as values in memory (estimated size 11.4 KB, free 365.9 MB)

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_3\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.9 MB)

23/05/26 07:19:27 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:27 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (PythonRDD[7] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:27 INFO cluster.YarnScheduler: Adding task set 2.0 with 1 tasks

23/05/26 07:19:27 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:27 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:27 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 95 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:27 INFO cluster.YarnScheduler: Removed TaskSet 2.0, whose tasks have all completed, from pool

23/05/26 07:19:27 INFO scheduler.DAGScheduler: ResultStage 2 (runJob at PythonRDD.scala:153) finished in 0.110 s

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Job 2 finished: runJob at PythonRDD.scala:153, took 0.114986 s

[1, 1, 1]

----> Iteration took: 1.9311904907226562e-05 ms

23/05/26 07:19:27 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Got job 3 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Final stage: ResultStage 3 (runJob at PythonRDD.scala:153)

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Submitting ResultStage 3 (PythonRDD[8] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_4 stored as values in memory (estimated size 11.4 KB, free 365.9 MB)

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_4\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.9 MB)

23/05/26 07:19:27 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:27 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:27 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 3 (PythonRDD[8] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:27 INFO cluster.YarnScheduler: Adding task set 3.0 with 1 tasks

23/05/26 07:19:27 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:27 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 91 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:28 INFO cluster.YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool

23/05/26 07:19:28 INFO scheduler.DAGScheduler: ResultStage 3 (runJob at PythonRDD.scala:153) finished in 0.103 s

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Job 3 finished: runJob at PythonRDD.scala:153, took 0.108842 s

[1, 1, 1]

----> Iteration took: 3.0517578125e-05 ms

23/05/26 07:19:28 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Got job 4 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Final stage: ResultStage 4 (runJob at PythonRDD.scala:153)

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting ResultStage 4 (PythonRDD[9] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_5 stored as values in memory (estimated size 11.4 KB, free 365.9 MB)

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_5\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.9 MB)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_5\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO spark.SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 4 (PythonRDD[9] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:28 INFO cluster.YarnScheduler: Adding task set 4.0 with 1 tasks

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 4.0 (TID 4, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_5\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 4.0 (TID 4) in 87 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:28 INFO cluster.YarnScheduler: Removed TaskSet 4.0, whose tasks have all completed, from pool

23/05/26 07:19:28 INFO scheduler.DAGScheduler: ResultStage 4 (runJob at PythonRDD.scala:153) finished in 0.098 s

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Job 4 finished: runJob at PythonRDD.scala:153, took 0.102163 s

[1, 1, 1]

----> Iteration took: 2.2649765014648438e-05 ms

23/05/26 07:19:28 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Got job 5 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Final stage: ResultStage 5 (runJob at PythonRDD.scala:153)

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting ResultStage 5 (PythonRDD[10] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_6 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_6\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_6\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO spark.SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (PythonRDD[10] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:28 INFO cluster.YarnScheduler: Adding task set 5.0 with 1 tasks

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 5.0 (TID 5, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_6\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 5.0 (TID 5) in 82 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:28 INFO cluster.YarnScheduler: Removed TaskSet 5.0, whose tasks have all completed, from pool

23/05/26 07:19:28 INFO scheduler.DAGScheduler: ResultStage 5 (runJob at PythonRDD.scala:153) finished in 0.094 s

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Job 5 finished: runJob at PythonRDD.scala:153, took 0.097783 s

[1, 1, 1]

----> Iteration took: 1.9788742065429688e-05 ms

23/05/26 07:19:28 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Got job 6 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Final stage: ResultStage 6 (runJob at PythonRDD.scala:153)

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting ResultStage 6 (PythonRDD[11] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_7 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_7\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_7\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO spark.SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 6 (PythonRDD[11] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:28 INFO cluster.YarnScheduler: Adding task set 6.0 with 1 tasks

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 6.0 (TID 6, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_7\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 6.0 (TID 6) in 79 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:28 INFO cluster.YarnScheduler: Removed TaskSet 6.0, whose tasks have all completed, from pool

23/05/26 07:19:28 INFO scheduler.DAGScheduler: ResultStage 6 (runJob at PythonRDD.scala:153) finished in 0.091 s

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Job 6 finished: runJob at PythonRDD.scala:153, took 0.096016 s

[1, 1, 1]

----> Iteration took: 2.1696090698242188e-05 ms

23/05/26 07:19:28 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Got job 7 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Final stage: ResultStage 7 (runJob at PythonRDD.scala:153)

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting ResultStage 7 (PythonRDD[12] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_8 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_8\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_8\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO spark.SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 7 (PythonRDD[12] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:28 INFO cluster.YarnScheduler: Adding task set 7.0 with 1 tasks

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 7.0 (TID 7, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_8\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 7.0 (TID 7) in 81 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:28 INFO cluster.YarnScheduler: Removed TaskSet 7.0, whose tasks have all completed, from pool

23/05/26 07:19:28 INFO scheduler.DAGScheduler: ResultStage 7 (runJob at PythonRDD.scala:153) finished in 0.095 s

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Job 7 finished: runJob at PythonRDD.scala:153, took 0.099005 s

[1, 1, 1]

----> Iteration took: 2.0265579223632812e-05 ms

23/05/26 07:19:28 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Got job 8 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Final stage: ResultStage 8 (runJob at PythonRDD.scala:153)

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting ResultStage 8 (PythonRDD[13] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_9 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_9\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_9\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO spark.SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 8 (PythonRDD[13] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:28 INFO cluster.YarnScheduler: Adding task set 8.0 with 1 tasks

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 8.0 (TID 8, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_9\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 8.0 (TID 8) in 82 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:28 INFO cluster.YarnScheduler: Removed TaskSet 8.0, whose tasks have all completed, from pool

23/05/26 07:19:28 INFO scheduler.DAGScheduler: ResultStage 8 (runJob at PythonRDD.scala:153) finished in 0.093 s

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Job 8 finished: runJob at PythonRDD.scala:153, took 0.098470 s

[1, 1, 1]

----> Iteration took: 1.9311904907226562e-05 ms

23/05/26 07:19:28 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:153

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Got job 9 (runJob at PythonRDD.scala:153) with 1 output partitions

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Final stage: ResultStage 9 (runJob at PythonRDD.scala:153)

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Parents of final stage: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Missing parents: List()

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting ResultStage 9 (PythonRDD[14] at RDD at PythonRDD.scala:53), which has no missing parents

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_10 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_10\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_10\_piece0 in memory on ip-10-1-1-204.ap-south-1.compute.internal:45048 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO spark.SparkContext: Created broadcast 10 from broadcast at DAGScheduler.scala:1164

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 9 (PythonRDD[14] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions Vector(0))

23/05/26 07:19:28 INFO cluster.YarnScheduler: Adding task set 9.0 with 1 tasks

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 9.0 (TID 9, ip-10-1-2-103.ap-south-1.compute.internal, executor 1, partition 0, PROCESS\_LOCAL, 8347 bytes)

23/05/26 07:19:28 INFO storage.BlockManagerInfo: Added broadcast\_10\_piece0 in memory on ip-10-1-2-103.ap-south-1.compute.internal:34192 (size: 6.3 KB, free: 366.2 MB)

23/05/26 07:19:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 9.0 (TID 9) in 74 ms on ip-10-1-2-103.ap-south-1.compute.internal (executor 1) (1/1)

23/05/26 07:19:28 INFO cluster.YarnScheduler: Removed TaskSet 9.0, whose tasks have all completed, from pool

23/05/26 07:19:28 INFO scheduler.DAGScheduler: ResultStage 9 (runJob at PythonRDD.scala:153) finished in 0.089 s

23/05/26 07:19:28 INFO scheduler.DAGScheduler: Job 9 finished: runJob at PythonRDD.scala:153, took 0.092472 s

[1, 1, 1]

23/05/26 07:19:28 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread

23/05/26 07:19:28 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors

23/05/26 07:19:28 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down

23/05/26 07:19:28 INFO cluster.SchedulerExtensionServices: Stopping SchedulerExtensionServices

(serviceOption=None,

services=List(),

started=false)

23/05/26 07:19:28 INFO cluster.YarnClientSchedulerBackend: Stopped

23/05/26 07:19:28 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!

23/05/26 07:19:28 INFO memory.MemoryStore: MemoryStore cleared

23/05/26 07:19:28 INFO storage.BlockManager: BlockManager stopped

23/05/26 07:19:28 INFO storage.BlockManagerMaster: BlockManagerMaster stopped

23/05/26 07:19:28 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

23/05/26 07:19:28 INFO spark.SparkContext: Successfully stopped SparkContext

23/05/26 07:19:29 INFO util.ShutdownHookManager: Shutdown hook called

23/05/26 07:19:29 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-6cdab5ed-bbba-4c2c-a7ec-441a4b64277d

23/05/26 07:19:29 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-6cdab5ed-bbba-4c2c-a7ec-441a4b64277d/pyspark-1cec93b0-6077-4d1f-a7e4-ff664c593c99

23/05/26 07:19:29 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-9de51f8e-6ede-422a-ba8b-069e20a14f6c

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

[raghavendr48edu@ip-10-1-1-204 ~]$ **yarn logs --applicationId application\_1684866872278\_2046**

WARNING: YARN\_OPTS has been replaced by HADOOP\_OPTS. Using value of YARN\_OPTS.

23/05/26 07:25:40 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

End of LogType:prelaunch.err

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Container: container\_1684866872278\_2046\_01\_000001 on ip-10-1-2-103.ap-south-1.compute.internal\_8041

LogAggregationType: AGGREGATED

===================================================================================================

LogType:prelaunch.out

LogLastModifiedTime:Fri May 26 07:19:30 +0000 2023

LogLength:70

LogContents:

Setting up env variables

Setting up job resources

Launching container

End of LogType:prelaunch.out

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Container: container\_1684866872278\_2046\_01\_000001 on ip-10-1-2-103.ap-south-1.compute.internal\_8041

LogAggregationType: AGGREGATED

===================================================================================================

LogType:stderr

LogLastModifiedTime:Fri May 26 07:19:30 +0000 2023

LogLength:34939

LogContents:

23/05/26 07:19:15 INFO util.SignalUtils: Registered signal handler for TERM

23/05/26 07:19:15 INFO util.SignalUtils: Registered signal handler for HUP

23/05/26 07:19:15 INFO util.SignalUtils: Registered signal handler for INT

23/05/26 07:19:15 INFO spark.SecurityManager: Changing view acls to: yarn,raghavendr48edu

23/05/26 07:19:15 INFO spark.SecurityManager: Changing modify acls to: yarn,raghavendr48edu

23/05/26 07:19:15 INFO spark.SecurityManager: Changing view acls groups to:

23/05/26 07:19:15 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/26 07:19:15 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(yarn, raghavendr48edu); groups with view permissions: Set(); users with modify permissions: Se

t(yarn, raghavendr48edu); groups with modify permissions: Set()

23/05/26 07:19:16 INFO yarn.ApplicationMaster: ApplicationAttemptId: appattempt\_1684866872278\_2046\_000001

23/05/26 07:19:16 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8030

23/05/26 07:19:16 INFO yarn.YarnRMClient: Registering the ApplicationMaster

23/05/26 07:19:17 INFO client.TransportClientFactory: Successfully created connection to ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:41949 after 84 ms (0 ms spent in bootstraps)

23/05/26 07:19:17 INFO yarn.ApplicationMaster: Preparing Local resources

23/05/26 07:19:17 INFO yarn.ApplicationMaster:

===============================================================================

YARN executor launch context:

env:

CLASSPATH -> {{PWD}}<CPS>{{PWD}}/\_\_spark\_conf\_\_<CPS>{{PWD}}/\_\_spark\_libs\_\_/\*<CPS>{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/jars/\*<CPS>{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/l

ib/spark/hive/\*<CPS>$HADOOP\_CLIENT\_CONF\_DIR<CPS>$HADOOP\_COMMON\_HOME/\*<CPS>$HADOOP\_COMMON\_HOME/lib/\*<CPS>$HADOOP\_HDFS\_HOME/\*<CPS>$HADOOP\_HDFS\_HOME/lib/\*<CPS>$HADOOP\_YARN\_HOME/\*<CPS>$HADOOP\_YARN\_HOME/lib/\*<CPS>$HADOOP\_CLIENT\_CONF\_DIR<CPS

>$PWD/mr-framework/\*<CPS>$MR2\_CLASSPATH<CPS>{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/accessors-smart-1.2.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/acc

essors-smart.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/asm-5.0.4.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/asm.jar:{{HADOOP\_COMMON\_HOME}}/../../../

CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/avro.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/aws-java-sdk-bundle-1.11.271.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.14257

74/lib/hadoop/client/aws-java-sdk-bundle.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/azure-data-lake-store-sdk-2.2.9.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hado

op/client/azure-data-lake-store-sdk.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-beanutils-1.9.3.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/com

mons-beanutils.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-cli-1.2.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-cli.jar:{{HADOOP\_COMMON\_

HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-codec-1.11.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-codec.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cd

h6.2.1.p0.1425774/lib/hadoop/client/commons-collections-3.2.2.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-collections.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774

/lib/hadoop/client/commons-compress-1.18.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-compress.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commo

ns-configuration2-2.1.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-configuration2.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-io-2.6.j

ar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-io.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-lang-2.6.jar:{{HADOOP\_COMMON\_HOME}}/../../../

CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-lang.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-lang3-3.7.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/

lib/hadoop/client/commons-lang3.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-logging-1.1.3.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-l

ogging.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-math3-3.1.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-math3.jar:{{HADOOP\_COMMON\_HO

ME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-net-3.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-net.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.

p0.1425774/lib/hadoop/client/curator-client-2.12.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-client.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/clie

nt/curator-framework-2.12.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-framework.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-recipes-2

.12.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-recipes.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/gson-2.2.4.jar:{{HADOOP\_COMMON\_HOME}}/../

../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/gson.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/guava-11.0.2.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/had

oop/client/guava.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-annotations-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-annot

ations.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-auth-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-auth.jar:{{HADOOP\_COMM

ON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-aws-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-aws.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2

.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-azure-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-azure-datalake-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6

.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-azure-datalake.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-azure.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/l

ib/hadoop/client/hadoop-common-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-common.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/had

oop-hdfs-client-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-hdfs-client.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapred

uce-client-common-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-common.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client

/hadoop-mapreduce-client-core-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-core.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/had

oop/client/hadoop-mapreduce-client-jobclient-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-jobclient.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.

1.p0.1425774/lib/hadoop/client/hadoop-yarn-api-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-api.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib

/hadoop/client/hadoop-yarn-client-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-client.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/cl

ient/hadoop-yarn-common-3.0.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-common.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/htrac

e-core4-4.1.0-incubating.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/htrace-core4.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/httpclient-4.5.3.jar:{{HA

DOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/httpclient.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/httpcore-4.4.6.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1

-1.cdh6.2.1.p0.1425774/lib/hadoop/client/httpcore.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jaxb-api-2.2.11.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/clie

nt/jaxb-api.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jcip-annotations-1.0-1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jcip-annotations.jar:{{HADOO

P\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/json-smart-2.3.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/json-smart.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.

cdh6.2.1.p0.1425774/lib/hadoop/client/jsp-api-2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsp-api.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsr30

5-3.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsr305.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsr311-api-1.1.1.jar:{{HADOOP\_COMMON\_HOME}}/../.

./../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsr311-api.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-admin-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.142577

4/lib/hadoop/client/kerb-admin.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-client-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-client.ja

r:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-common-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-common.jar:{{HADOOP\_COMMON\_HOME}}/../../..

/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-core-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-core.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/

hadoop/client/kerb-crypto-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-crypto.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-identity-1.0.0

.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-identity.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-server-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../

../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-server.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-simplekdc-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.

1425774/lib/hadoop/client/kerb-simplekdc.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-util-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-u

til.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-asn1-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-asn1.jar:{{HADOOP\_COMMON\_HOME}}/../.

./../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-config-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-config.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.14

25774/lib/hadoop/client/kerby-pkix-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-pkix.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-util-

1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-util.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-xdr-1.0.0.jar:{{HADOOP\_COMMON\_HOME}}/../

../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-xdr.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/log4j-1.2.17.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/li

b/hadoop/client/log4j.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/nimbus-jose-jwt-4.41.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/nimbus-jose-jwt.ja

r:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okhttp-2.7.5.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okhttp.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1

-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okio-1.6.0.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okio.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/parana

mer-2.8.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/paranamer.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/protobuf-java-2.5.0.jar:{{HADOOP\_COMMON\_HOME}

}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/protobuf-java.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/re2j-1.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.14257

74/lib/hadoop/client/re2j.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/snappy-java-1.1.4.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/snappy-java.jar:{{H

ADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/stax2-api-3.1.4.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/stax2-api.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.

1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/wildfly-openssl-1.0.4.Final.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/wildfly-openssl.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.142

5774/lib/hadoop/client/woodstox-core-5.0.3.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/woodstox-core.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/xz-1.6

.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/xz.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-thirdparty/audience-annotations-0.5.0.jar:

{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-thirdparty/commons-logging-1.2.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-th

irdparty/findbugs-annotations-1.3.9-1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-thirdparty/htrace-core4-4.2.0-incubating.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.

2.1.p0.1425774/lib/hbase/bin/../lib/shaded-clients/hbase-shaded-mapreduce-2.1.0-cdh6.2.1.jar:{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/hbase-spark.jar<CPS>{{PWD}}/\_\_spark\_conf\_\_/\_\_hadoop\_conf\_\_

MKL\_NUM\_THREADS -> 1

SPARK\_DIST\_CLASSPATH -> /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/accessors-smart-1.2.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/accessors-smart.jar:/opt/cloudera/parce

ls/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/asm-5.0.4.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/asm.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/avro.jar:/opt/cl

oudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/aws-java-sdk-bundle-1.11.271.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/aws-java-sdk-bundle.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.

1.p0.1425774/lib/hadoop/client/azure-data-lake-store-sdk-2.2.9.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/azure-data-lake-store-sdk.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/c

lient/commons-beanutils-1.9.3.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-beanutils.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-cli-1.2.jar:/opt/cloudera/p

arcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-cli.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-codec-1.11.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/cl

ient/commons-codec.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-collections-3.2.2.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-collections.jar:/opt/cloudera/

parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-compress-1.18.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-compress.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/

hadoop/client/commons-configuration2-2.1.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-configuration2.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-io-2.6.ja

r:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-io.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-lang-2.6.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774

/lib/hadoop/client/commons-lang.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-lang3-3.7.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-lang3.jar:/opt/cloudera/p

arcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-logging-1.1.3.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-logging.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/ha

doop/client/commons-math3-3.1.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-math3.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-net-3.1.jar:/opt/cloudera/par

cels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/commons-net.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-client-2.12.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/c

lient/curator-client.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-framework-2.12.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-framework.jar:/opt/cloudera/p

arcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-recipes-2.12.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/curator-recipes.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/h

adoop/client/gson-2.2.4.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/gson.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/guava-11.0.2.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh

6.2.1.p0.1425774/lib/hadoop/client/guava.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-annotations-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-a

nnotations.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-auth-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-auth.jar:/opt/cloudera/parcels/CDH-6.2

.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-aws-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-aws.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/had

oop-azure-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-azure-datalake-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-azure-datalake

.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-azure.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-common-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cd

h6.2.1.p0.1425774/lib/hadoop/client/hadoop-common.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-hdfs-client-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client

/hadoop-hdfs-client.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-common-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-

client-common.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-core-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-c

ore.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-jobclient-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-mapreduce-client-jobcli

ent.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-api-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-api.jar:/opt/cloudera/parcels/CDH-6.

2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-client-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-client.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/

hadoop/client/hadoop-yarn-common-3.0.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/hadoop-yarn-common.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/htrace-core4-4.1

.0-incubating.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/htrace-core4.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/httpclient-4.5.3.jar:/opt/cloudera/parcels/CDH-6.2.1-1.c

dh6.2.1.p0.1425774/lib/hadoop/client/httpclient.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/httpcore-4.4.6.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/httpcore.jar:/opt/cl

oudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jaxb-api-2.2.11.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jaxb-api.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/

client/jcip-annotations-1.0-1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jcip-annotations.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/json-smart-2.3.jar:/opt/cloudera/par

cels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/json-smart.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsp-api-2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsp-a

pi.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsr305-3.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsr305.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/li

b/hadoop/client/jsr311-api-1.1.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/jsr311-api.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-admin-1.0.0.jar:/opt/cloudera/parc

els/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-admin.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-client-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/

kerb-client.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-common-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-common.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh

6.2.1.p0.1425774/lib/hadoop/client/kerb-core-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-core.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-crypto-1.0.0.jar:

/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-crypto.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-identity-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.14257

74/lib/hadoop/client/kerb-identity.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-server-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-server.jar:/opt/cloudera/

parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-simplekdc-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-simplekdc.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/had

oop/client/kerb-util-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerb-util.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-asn1-1.0.0.jar:/opt/cloudera/parcels/CDH

-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-asn1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-config-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-

config.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-pkix-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-pkix.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p

0.1425774/lib/hadoop/client/kerby-util-1.0.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-util.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-xdr-1.0.0.jar:/opt/cl

oudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/kerby-xdr.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/log4j-1.2.17.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/cl

ient/log4j.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/nimbus-jose-jwt-4.41.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/nimbus-jose-jwt.jar:/opt/cloudera/parcels/CDH-6.2

.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okhttp-2.7.5.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okhttp.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okio-1.6.0.jar:/opt/

cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/okio.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/paranamer-2.8.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/clie

nt/paranamer.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/protobuf-java-2.5.0.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/protobuf-java.jar:/opt/cloudera/parcels/CDH-6.2.1-

1.cdh6.2.1.p0.1425774/lib/hadoop/client/re2j-1.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/re2j.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/snappy-java-1.1.4.jar:/opt/cl

oudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/snappy-java.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/stax2-api-3.1.4.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hado

op/client/stax2-api.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/wildfly-openssl-1.0.4.Final.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/wildfly-openssl.jar:/opt/cloudera/p

arcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/woodstox-core-5.0.3.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/woodstox-core.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop

/client/xz-1.6.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/client/xz.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-thirdparty/audience-annotations-0.5.0.jar:/opt/c

loudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-thirdparty/commons-logging-1.2.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-thirdparty/findbugs-annotati

ons-1.3.9-1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/client-facing-thirdparty/htrace-core4-4.2.0-incubating.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/bin/../lib/shaded-cli

ents/hbase-shaded-mapreduce-2.1.0-cdh6.2.1.jar:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hbase/hbase-spark.jar

SPARK\_YARN\_STAGING\_DIR -> hdfs://nameservice1/user/raghavendr48edu/.sparkStaging/application\_1684866872278\_2046

SPARK\_USER -> raghavendr48edu

PYTHONPATH -> /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/python/lib/py4j-0.10.7-src.zip<CPS>/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/python/lib/pyspark.zip

OPENBLAS\_NUM\_THREADS -> 1

command:

LD\_LIBRARY\_PATH=\"{{HADOOP\_COMMON\_HOME}}/../../../CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/lib/native:$LD\_LIBRARY\_PATH\" \

{{JAVA\_HOME}}/bin/java \

-server \

-Xmx1024m \

-Djava.io.tmpdir={{PWD}}/tmp \

'-Dspark.authenticate=false' \

'-Dspark.driver.port=41949' \

'-Dspark.network.crypto.enabled=false' \

'-Dspark.shuffle.service.port=7337' \

-Dspark.yarn.app.container.log.dir=<LOG\_DIR> \

-XX:OnOutOfMemoryError='kill %p' \

org.apache.spark.executor.CoarseGrainedExecutorBackend \

--driver-url \

spark://CoarseGrainedScheduler@ip-10-1-1-204.ap-south-1.compute.internal:41949 \

--executor-id \

<executorId> \

--hostname \

<hostname> \

--cores \

1 \

--app-id \

application\_1684866872278\_2046 \

--user-class-path \

[file:$PWD/\_\_app\_\_.jar](file:///\\$PWD\__app__.jar) \

1><LOG\_DIR>/stdout \

2><LOG\_DIR>/stderr

resources:

\_\_spark\_conf\_\_ -> resource { scheme: "hdfs" host: "nameservice1" port: -1 file: "/user/raghavendr48edu/.sparkStaging/application\_1684866872278\_2046/\_\_spark\_conf\_\_.zip" } size: 169649 timestamp: 1685085554127 type: ARCHIVE visibilit

y: PRIVATE

===============================================================================

23/05/26 07:19:17 INFO util.Utils: Using initial executors = 0, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances

23/05/26 07:19:17 INFO conf.Configuration: resource-types.xml not found

23/05/26 07:19:17 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/26 07:19:18 INFO yarn.ApplicationMaster: Started progress reporter thread with (heartbeat : 3000, initial allocation : 200) intervals

23/05/26 07:19:21 INFO yarn.YarnAllocator: Driver requested a total number of 1 executor(s).

23/05/26 07:19:21 INFO yarn.YarnAllocator: Will request 1 executor container(s), each with 1 core(s) and 1408 MB memory (including 384 MB of overhead)

23/05/26 07:19:21 INFO yarn.YarnAllocator: Submitted container request for host ip-10-1-2-40.ap-south-1.compute.internal,ip-10-1-2-175.ap-south-1.compute.internal,ip-10-1-2-103.ap-south-1.compute.internal.

23/05/26 07:19:21 INFO yarn.YarnAllocator: Launching container container\_1684866872278\_2046\_01\_000002 on host ip-10-1-2-103.ap-south-1.compute.internal for executor with ID 1

23/05/26 07:19:21 INFO yarn.YarnAllocator: Received 1 containers from YARN, launching executors on 1 of them.

23/05/26 07:19:27 INFO yarn.YarnAllocator: Driver requested a total number of 0 executor(s).

23/05/26 07:19:28 INFO yarn.ApplicationMaster$AMEndpoint: Driver terminated or disconnected! Shutting down. ip-10-1-1-204.ap-south-1.compute.internal:41949

23/05/26 07:19:28 INFO yarn.ApplicationMaster$AMEndpoint: Driver terminated or disconnected! Shutting down. ip-10-1-1-204.ap-south-1.compute.internal:41949

23/05/26 07:19:28 INFO yarn.ApplicationMaster: Final app status: SUCCEEDED, exitCode: 0

23/05/26 07:19:28 INFO yarn.ApplicationMaster: Unregistering ApplicationMaster with SUCCEEDED

23/05/26 07:19:28 INFO impl.AMRMClientImpl: Waiting for application to be successfully unregistered.

23/05/26 07:19:28 INFO yarn.ApplicationMaster: Deleting staging directory hdfs://nameservice1/user/raghavendr48edu/.sparkStaging/application\_1684866872278\_2046

23/05/26 07:19:28 INFO util.ShutdownHookManager: Shutdown hook called

End of LogType:stderr

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

End of LogType:stdout

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

End of LogType:prelaunch.err

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Container: container\_1684866872278\_2046\_01\_000002 on ip-10-1-2-103.ap-south-1.compute.internal\_8041

LogAggregationType: AGGREGATED

===================================================================================================

LogType:prelaunch.out

LogLastModifiedTime:Fri May 26 07:19:30 +0000 2023

LogLength:70

LogContents:

Setting up env variables

Setting up job resources

Launching container

End of LogType:prelaunch.out

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Container: container\_1684866872278\_2046\_01\_000002 on ip-10-1-2-103.ap-south-1.compute.internal\_8041

LogAggregationType: AGGREGATED

===================================================================================================

LogType:stderr

LogLastModifiedTime:Fri May 26 07:19:30 +0000 2023

LogLength:14170

LogContents:

23/05/26 07:19:22 INFO executor.CoarseGrainedExecutorBackend: Started daemon with process name: 4899@ip-10-1-2-103.ap-south-1.compute.internal

23/05/26 07:19:22 INFO util.SignalUtils: Registered signal handler for TERM

23/05/26 07:19:22 INFO util.SignalUtils: Registered signal handler for HUP

23/05/26 07:19:22 INFO util.SignalUtils: Registered signal handler for INT

23/05/26 07:19:23 INFO spark.SecurityManager: Changing view acls to: yarn,raghavendr48edu

23/05/26 07:19:23 INFO spark.SecurityManager: Changing modify acls to: yarn,raghavendr48edu

23/05/26 07:19:23 INFO spark.SecurityManager: Changing view acls groups to:

23/05/26 07:19:23 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/26 07:19:23 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(yarn, raghavendr48edu); groups with view permissions: Set(); users with modify permissions: Se

t(yarn, raghavendr48edu); groups with modify permissions: Set()

23/05/26 07:19:23 INFO client.TransportClientFactory: Successfully created connection to ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:41949 after 89 ms (0 ms spent in bootstraps)

23/05/26 07:19:23 INFO spark.SecurityManager: Changing view acls to: yarn,raghavendr48edu

23/05/26 07:19:23 INFO spark.SecurityManager: Changing modify acls to: yarn,raghavendr48edu

23/05/26 07:19:23 INFO spark.SecurityManager: Changing view acls groups to:

23/05/26 07:19:23 INFO spark.SecurityManager: Changing modify acls groups to:

23/05/26 07:19:23 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(yarn, raghavendr48edu); groups with view permissions: Set(); users with modify permissions: Se

t(yarn, raghavendr48edu); groups with modify permissions: Set()

23/05/26 07:19:23 INFO client.TransportClientFactory: Successfully created connection to ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:41949 after 1 ms (0 ms spent in bootstraps)

23/05/26 07:19:23 INFO storage.DiskBlockManager: Created local directory at /mnt/data/yarn/nm/usercache/raghavendr48edu/appcache/application\_1684866872278\_2046/blockmgr-bf18a20f-59ab-45e1-a4a5-79ad49c37cea

23/05/26 07:19:23 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB

23/05/26 07:19:24 INFO executor.CoarseGrainedExecutorBackend: Connecting to driver: spark://CoarseGrainedScheduler@ip-10-1-1-204.ap-south-1.compute.internal:41949

23/05/26 07:19:24 INFO executor.CoarseGrainedExecutorBackend: Successfully registered with driver

23/05/26 07:19:24 INFO executor.Executor: Starting executor ID 1 on host ip-10-1-2-103.ap-south-1.compute.internal

23/05/26 07:19:24 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 34192.

23/05/26 07:19:24 INFO netty.NettyBlockTransferService: Server created on ip-10-1-2-103.ap-south-1.compute.internal:34192

23/05/26 07:19:24 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy

23/05/26 07:19:24 INFO storage.BlockManagerMaster: Registering BlockManager BlockManagerId(1, ip-10-1-2-103.ap-south-1.compute.internal, 34192, None)

23/05/26 07:19:24 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(1, ip-10-1-2-103.ap-south-1.compute.internal, 34192, None)

23/05/26 07:19:24 INFO storage.BlockManager: external shuffle service port = 7337

23/05/26 07:19:24 INFO storage.BlockManager: Registering executor with local external shuffle service.

23/05/26 07:19:24 INFO client.TransportClientFactory: Successfully created connection to ip-10-1-2-103.ap-south-1.compute.internal/10.1.2.103:7337 after 1 ms (0 ms spent in bootstraps)

23/05/26 07:19:24 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(1, ip-10-1-2-103.ap-south-1.compute.internal, 34192, None)

23/05/26 07:19:24 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 0

23/05/26 07:19:24 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)

23/05/26 07:19:24 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 1

23/05/26 07:19:24 INFO client.TransportClientFactory: Successfully created connection to ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:45048 after 1 ms (0 ms spent in bootstraps)

23/05/26 07:19:24 INFO memory.MemoryStore: Block broadcast\_1\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 366.3 MB)

23/05/26 07:19:24 INFO broadcast.TorrentBroadcast: Reading broadcast variable 1 took 125 ms

23/05/26 07:19:24 INFO memory.MemoryStore: Block broadcast\_1 stored as values in memory (estimated size 11.4 KB, free 366.3 MB)

23/05/26 07:19:25 INFO codegen.CodeGenerator: Code generated in 280.931924 ms

23/05/26 07:19:26 INFO datasources.FileScanRDD: Reading File path: hdfs://nameservice1/user/raghavendr48edu/input\_sort\_py.txt, range: 0-85, partition values: [empty row]

23/05/26 07:19:26 INFO codegen.CodeGenerator: Code generated in 18.885048 ms

23/05/26 07:19:26 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 0

23/05/26 07:19:26 INFO memory.MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 32.3 KB, free 366.3 MB)

23/05/26 07:19:26 INFO broadcast.TorrentBroadcast: Reading broadcast variable 0 took 14 ms

23/05/26 07:19:26 INFO memory.MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 443.5 KB, free 365.8 MB)

23/05/26 07:19:27 INFO python.PythonRunner: Times: total = 1850, boot = 794, init = 1055, finish = 1

23/05/26 07:19:27 INFO memory.MemoryStore: Block rdd\_4\_0 stored as values in memory (estimated size 69.0 B, free 365.8 MB)

23/05/26 07:19:27 INFO python.PythonRunner: Times: total = 42, boot = -19, init = 60, finish = 1

23/05/26 07:19:27 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 1886 bytes result sent to driver

23/05/26 07:19:27 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 1

23/05/26 07:19:27 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)

23/05/26 07:19:27 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 2

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_2\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:27 INFO broadcast.TorrentBroadcast: Reading broadcast variable 2 took 14 ms

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_2 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:27 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:27 INFO python.PythonRunner: Times: total = 40, boot = -182, init = 222, finish = 0

23/05/26 07:19:27 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 1800 bytes result sent to driver

23/05/26 07:19:27 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 2

23/05/26 07:19:27 INFO executor.Executor: Running task 0.0 in stage 2.0 (TID 2)

23/05/26 07:19:27 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 3

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_3\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:27 INFO broadcast.TorrentBroadcast: Reading broadcast variable 3 took 13 ms

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_3 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:27 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:27 INFO python.PythonRunner: Times: total = 40, boot = -88, init = 128, finish = 0

23/05/26 07:19:27 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 2). 1800 bytes result sent to driver

23/05/26 07:19:27 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 3

23/05/26 07:19:27 INFO executor.Executor: Running task 0.0 in stage 3.0 (TID 3)

23/05/26 07:19:27 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 4

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_4\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:27 INFO broadcast.TorrentBroadcast: Reading broadcast variable 4 took 14 ms

23/05/26 07:19:27 INFO memory.MemoryStore: Block broadcast\_4 stored as values in memory (estimated size 11.4 KB, free 365.8 MB)

23/05/26 07:19:27 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:27 INFO python.PythonRunner: Times: total = 41, boot = -71, init = 112, finish = 0

23/05/26 07:19:27 INFO executor.Executor: Finished task 0.0 in stage 3.0 (TID 3). 1800 bytes result sent to driver

23/05/26 07:19:28 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 4

23/05/26 07:19:28 INFO executor.Executor: Running task 0.0 in stage 4.0 (TID 4)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 5

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_5\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.8 MB)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Reading broadcast variable 5 took 13 ms

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_5 stored as values in memory (estimated size 11.4 KB, free 365.7 MB)

23/05/26 07:19:28 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:28 INFO python.PythonRunner: Times: total = 40, boot = -76, init = 116, finish = 0

23/05/26 07:19:28 INFO executor.Executor: Finished task 0.0 in stage 4.0 (TID 4). 1800 bytes result sent to driver

23/05/26 07:19:28 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 5

23/05/26 07:19:28 INFO executor.Executor: Running task 0.0 in stage 5.0 (TID 5)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 6

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_6\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.7 MB)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Reading broadcast variable 6 took 13 ms

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_6 stored as values in memory (estimated size 11.4 KB, free 365.7 MB)

23/05/26 07:19:28 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:28 INFO python.PythonRunner: Times: total = 41, boot = -64, init = 105, finish = 0

23/05/26 07:19:28 INFO executor.Executor: Finished task 0.0 in stage 5.0 (TID 5). 1843 bytes result sent to driver

23/05/26 07:19:28 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 6

23/05/26 07:19:28 INFO executor.Executor: Running task 0.0 in stage 6.0 (TID 6)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 7

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_7\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.7 MB)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Reading broadcast variable 7 took 11 ms

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_7 stored as values in memory (estimated size 11.4 KB, free 365.7 MB)

23/05/26 07:19:28 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:28 INFO python.PythonRunner: Times: total = 42, boot = -60, init = 102, finish = 0

23/05/26 07:19:28 INFO executor.Executor: Finished task 0.0 in stage 6.0 (TID 6). 1800 bytes result sent to driver

23/05/26 07:19:28 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 7

23/05/26 07:19:28 INFO executor.Executor: Running task 0.0 in stage 7.0 (TID 7)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 8

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_8\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.7 MB)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Reading broadcast variable 8 took 11 ms

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_8 stored as values in memory (estimated size 11.4 KB, free 365.7 MB)

23/05/26 07:19:28 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:28 INFO python.PythonRunner: Times: total = 40, boot = -60, init = 100, finish = 0

23/05/26 07:19:28 INFO executor.Executor: Finished task 0.0 in stage 7.0 (TID 7). 1800 bytes result sent to driver

23/05/26 07:19:28 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 8

23/05/26 07:19:28 INFO executor.Executor: Running task 0.0 in stage 8.0 (TID 8)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 9

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_9\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.7 MB)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Reading broadcast variable 9 took 11 ms

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_9 stored as values in memory (estimated size 11.4 KB, free 365.7 MB)

23/05/26 07:19:28 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:28 INFO python.PythonRunner: Times: total = 41, boot = -63, init = 104, finish = 0

23/05/26 07:19:28 INFO executor.Executor: Finished task 0.0 in stage 8.0 (TID 8). 1800 bytes result sent to driver

23/05/26 07:19:28 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 9

23/05/26 07:19:28 INFO executor.Executor: Running task 0.0 in stage 9.0 (TID 9)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Started reading broadcast variable 10

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_10\_piece0 stored as bytes in memory (estimated size 6.3 KB, free 365.7 MB)

23/05/26 07:19:28 INFO broadcast.TorrentBroadcast: Reading broadcast variable 10 took 10 ms

23/05/26 07:19:28 INFO memory.MemoryStore: Block broadcast\_10 stored as values in memory (estimated size 11.4 KB, free 365.7 MB)

23/05/26 07:19:28 INFO storage.BlockManager: Found block rdd\_4\_0 locally

23/05/26 07:19:28 INFO python.PythonRunner: Times: total = 41, boot = -64, init = 105, finish = 0

23/05/26 07:19:28 INFO executor.Executor: Finished task 0.0 in stage 9.0 (TID 9). 1800 bytes result sent to driver

23/05/26 07:19:28 INFO executor.CoarseGrainedExecutorBackend: Driver commanded a shutdown

23/05/26 07:19:28 INFO memory.MemoryStore: MemoryStore cleared

23/05/26 07:19:28 INFO storage.BlockManager: BlockManager stopped

23/05/26 07:19:28 INFO util.ShutdownHookManager: Shutdown hook called

End of LogType:stderr

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

End of LogType:stdout

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

6.Transfer the sample dataset from RDBMS to HDFS

**Step 9.1: Download Dataset -Download dataset and transfer to cloud lab using FTP**

**Step 9.2: Load data in mysql - Login into mysql - Create schema in mysql**

show databases;

use database

Show tables;

127 login: raghavendr48edu

raghavendr48edu@127.0.0.1's password:

Last login: Thu May 18 14:47:35 2023 from localhost

[raghavendr48edu@ip-10-1-1-204 ~]$ **mysql -u raghavendr48edu -pRedKangaroo99$**

mysql: [Warning] Using a password on the command line interface can be insecure.

Welcome to the MySQL monitor. Commands end with ; or \g.

Your MySQL connection id is 3441

Server version: 5.7.33 MySQL Community Server (GPL)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its

affiliates. Other names may be trademarks of their respective

owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

Or

[raghavendr48edu@ip-10-1-1-204 ~]$ **mysql -u raghavendr48edu -p**

Enter password:

Welcome to the MySQL monitor. Commands end with ; or \g.

Your MySQL connection id is 3538

Server version: 5.7.33 MySQL Community Server (GPL)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its

affiliates. Other names may be trademarks of their respective

owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> **select @@hostname;**

+-------------------------------------------+

| @@hostname |

+-------------------------------------------+

| ip-10-1-1-204.ap-south-1.compute.internal |

+-------------------------------------------+

1 row in set (0.00 sec)

mysql> **create table**

mysql> show databases

-> show databases;

ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'sh

ow databases' at line 2

mysql> **show databases;**

+--------------------+

| Database |

+--------------------+

| information\_schema |

| raghavendr48edu |

+--------------------+

2 rows in set (0.03 sec)

mysql> **use raghavendr48edu**

Database changed

- **Load csv in table**

mysql> create table aisles(aisle\_id int(10),aisle varchar(100));

Query OK, 0 rows affected (0.01 sec)

mysql> load data local infile '/home/raghavendr48edu/aisles.csv' into table aisles FIELDS TERMINATED BY ',' IGNORE 1 ROWS;

Query OK, 134 rows affected (0.00 sec)

Records: 134 Deleted: 0 Skipped: 0 Warnings: 0

mysql> create table products(product\_id int(11),product\_name varchar(100),aisle\_id int(10),department\_id int(10));

Query OK, 0 rows affected (0.01 sec)

mysql> load data local infile '/home/raghavendr48edu/products.csv' into table products FIELDS TERMINATED BY ',' IGNORE 1 ROWS;

Query OK, 49688 rows affected, 7743 warnings (0.29 sec)

Records: 49688 Deleted: 0 Skipped: 0 Warnings: 7743

mysql> create table product\_aisles as select \* from products p join aisles a on p.aisle\_id=a.aisle\_id;

ERROR 1060 (42S21): Duplicate column name 'aisle\_id'

mysql> create table product\_aisles as select p.product\_id,p.product\_name,p.aisle\_id,p.department\_id,a.aisle from products p join aisles a on p.aisle\_id=a.aisle\_id;

Query OK, 46626 rows affected (0.83 sec)

Records: 46626 Duplicates: 0 Warnings: 0

**Step 9.3: Transferring data from Mysql to HDFS - List Databases**

[raghavendr48edu@ip-10-1-1-204 ~]$ **sqoop import --connect jdbc:mysql://ip-10-1-1-204.ap-south-1.compute.internal:3306/raghavendr48edu --username raghavendr48edu --password RedKangaroo99$ --table product\_aisles**

**--m 1 --driver com.mysql.jdbc.Driver --target-dir /user/raghavendr48edu/product\_aisles\_data/**

Warning: /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.

Please set $ACCUMULO\_HOME to the root of your Accumulo installation.

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See  <http://www.slf4j.org/codes.html#multiple_bindings> for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

23/05/21 13:15:37 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7-cdh6.2.1

23/05/21 13:15:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.

23/05/21 13:15:38 WARN sqoop.ConnFactory: Parameter --driver is set to an explicit driver however appropriate connection manager is not being set (via --connection-manager). Sqoop is going to fall back to org.

apache.sqoop.manager.GenericJdbcManager. Please specify explicitly which connection manager should be used next time.

23/05/21 13:15:38 INFO manager.SqlManager: Using default fetchSize of 1000

23/05/21 13:15:38 INFO tool.CodeGenTool: Beginning code generation

23/05/21 13:15:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 13:15:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 13:15:38 INFO orm.CompilationManager: HADOOP\_MAPRED\_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce

23/05/21 13:15:40 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-raghavendr48edu/compile/f7fc58672e3f4044b68b01821200b3a5/product\_aisles.jar

23/05/21 13:15:40 INFO mapreduce.ImportJobBase: Beginning import of product\_aisles

23/05/21 13:15:40 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar

23/05/21 13:15:40 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 13:15:41 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps

23/05/21 13:15:41 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/21 13:15:41 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/raghavendr48edu/.staging/job\_1684298513961\_0934

23/05/21 13:15:44 INFO db.DBInputFormat: Using read commited transaction isolation

23/05/21 13:15:44 INFO mapreduce.JobSubmitter: number of splits:1

23/05/21 13:15:44 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled

23/05/21 13:15:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1684298513961\_0934

23/05/21 13:15:44 INFO mapreduce.JobSubmitter: Executing with tokens: []

23/05/21 13:15:44 INFO conf.Configuration: resource-types.xml not found

23/05/21 13:15:44 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/21 13:15:44 INFO impl.YarnClientImpl: Submitted application application\_1684298513961\_0934

23/05/21 13:15:44 INFO mapreduce.Job: The url to track the job:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_0934/>

23/05/21 13:15:44 INFO mapreduce.Job: Running job: job\_1684298513961\_0934

23/05/21 13:15:52 INFO mapreduce.Job: Job job\_1684298513961\_0934 running in uber mode : false

23/05/21 13:15:52 INFO mapreduce.Job: map 0% reduce 0%

23/05/21 13:16:00 INFO mapreduce.Job: map 100% reduce 0%

23/05/21 13:16:00 INFO mapreduce.Job: Job job\_1684298513961\_0934 completed successfully

23/05/21 13:16:00 INFO mapreduce.Job: Counters: 33

File System Counters

FILE: Number of bytes read=0

FILE: Number of bytes written=246445

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=85

HDFS: Number of bytes written=2716265

HDFS: Number of read operations=6

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

HDFS: Number of bytes read erasure-coded=0

Job Counters

Launched map tasks=1

Other local map tasks=1

Total time spent by all maps in occupied slots (ms)=4150

Total time spent by all reduces in occupied slots (ms)=0

Total time spent by all map tasks (ms)=4150

Total vcore-milliseconds taken by all map tasks=4150

Total megabyte-milliseconds taken by all map tasks=4249600

Map-Reduce Framework

Map input records=46626

Map output records=46626

Input split bytes=85

Spilled Records=0

Failed Shuffles=0

Merged Map outputs=0

GC time elapsed (ms)=108

CPU time spent (ms)=2540

Physical memory (bytes) snapshot=384143360

Virtual memory (bytes) snapshot=2597126144

Total committed heap usage (bytes)=480247808

Peak Map Physical memory (bytes)=384143360

Peak Map Virtual memory (bytes)=2597126144

File Input Format Counters

Bytes Read=0

File Output Format Counters

Bytes Written=2716265

23/05/21 13:16:00 INFO mapreduce.ImportJobBase: Transferred 2.5904 MB in 19.126 seconds (138.6912 KB/sec)

23/05/21 13:16:00 INFO mapreduce.ImportJobBase: Retrieved 46626 records.

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

7.Validate the loaded data by comparing the statistics of data both in source

and HDFS

[raghavendr48edu@ip-10-1-1-204 ~]$ **hdfs dfs -cat /user/raghavendr48edu/product\_aisles\_data/\*|wc -l**

46626

[raghavendr48edu@ip-10-1-1-204 ~]$ **hdfs dfs -cat /user/raghavendr48edu/product\_aisles\_data/\*|head -5**

1,Chocolate Sandwich Cookies,61,19,cookies cakes

2,All-Seasons Salt,104,13,spices seasonings

3,Robust Golden Unsweetened Oolong Tea,94,7,tea

4,Smart Ones Classic Favorites Mini Rigatoni With Vodka Cream Sauce,38,1,frozen meals

5,Green Chile Anytime Sauce,5,13,marinades meat preparation

cat: Unable to write to output stream.

8.Create a new directory in HDFS named cheeses and load only rows where

aisle is “specialty cheeses“

[raghavendr48edu@ip-10-1-1-204 ~]$ **sqoop import --connect jdbc:mysql://ip-10-1-1-204.ap-south-1.compute.internal:3306/raghavendr48edu --username raghavendr48edu --password RedKangaroo99$ -m 1 --table product\_aisles --where "aisle = 'specialty cheeses'" --target-dir /user/raghavendr48edu/product\_aisle\_data/cheeses --driver com.mysql.jdbc.Driver**

Warning: /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.

Please set $ACCUMULO\_HOME to the root of your Accumulo installation.

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See  <http://www.slf4j.org/codes.html#multiple_bindings> for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

23/05/21 13:56:57 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7-cdh6.2.1

23/05/21 13:56:57 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.

23/05/21 13:56:57 WARN sqoop.ConnFactory: Parameter --driver is set to an explicit driver however appropriate connection manager is not being set (via --connection-manager). Sqoop is going to fall back to org.

apache.sqoop.manager.GenericJdbcManager. Please specify explicitly which connection manager should be used next time.

23/05/21 13:56:57 INFO manager.SqlManager: Using default fetchSize of 1000

23/05/21 13:56:57 INFO tool.CodeGenTool: Beginning code generation

23/05/21 13:56:58 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 13:56:58 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 13:56:58 INFO orm.CompilationManager: HADOOP\_MAPRED\_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce

23/05/21 13:56:59 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-raghavendr48edu/compile/f7bf53b4e3336cbc9e9f8d44198cc879/product\_aisles.java to /home/raghavendr48edu/./product\_aisles.java. Error: D

estination '/home/raghavendr48edu/./product\_aisles.java' already exists

23/05/21 13:56:59 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-raghavendr48edu/compile/f7bf53b4e3336cbc9e9f8d44198cc879/product\_aisles.jar

23/05/21 13:56:59 INFO mapreduce.ImportJobBase: Beginning import of product\_aisles

23/05/21 13:57:00 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar

23/05/21 13:57:00 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 13:57:00 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps

23/05/21 13:57:00 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/21 13:57:01 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/raghavendr48edu/.staging/job\_1684298513961\_0941

23/05/21 13:57:03 INFO db.DBInputFormat: Using read commited transaction isolation

23/05/21 13:57:03 INFO mapreduce.JobSubmitter: number of splits:1

23/05/21 13:57:03 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled

23/05/21 13:57:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1684298513961\_0941

23/05/21 13:57:03 INFO mapreduce.JobSubmitter: Executing with tokens: []

23/05/21 13:57:03 INFO conf.Configuration: resource-types.xml not found

23/05/21 13:57:03 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/21 13:57:03 INFO impl.YarnClientImpl: Submitted application application\_1684298513961\_0941

23/05/21 13:57:03 INFO mapreduce.Job: The url to track the job:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_0941/>

23/05/21 13:57:03 INFO mapreduce.Job: Running job: job\_1684298513961\_0941

23/05/21 13:57:10 INFO mapreduce.Job: Job job\_1684298513961\_0941 running in uber mode : false

23/05/21 13:57:10 INFO mapreduce.Job: map 0% reduce 0%

23/05/21 13:57:17 INFO mapreduce.Job: map 100% reduce 0%

23/05/21 13:57:17 INFO mapreduce.Job: Job job\_1684298513961\_0941 completed successfully

23/05/21 13:57:18 INFO mapreduce.Job: Counters: 33

File System Counters

FILE: Number of bytes read=0

FILE: Number of bytes written=246826

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=85

HDFS: Number of bytes written=14985

HDFS: Number of read operations=6

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

HDFS: Number of bytes read erasure-coded=0

Job Counters

Launched map tasks=1

Other local map tasks=1

Total time spent by all maps in occupied slots (ms)=4088

Total time spent by all reduces in occupied slots (ms)=0

Total time spent by all map tasks (ms)=4088

Total vcore-milliseconds taken by all map tasks=4088

Total megabyte-milliseconds taken by all map tasks=4186112

Map-Reduce Framework

Map input records=274

Map output records=274

Input split bytes=85

Spilled Records=0

Failed Shuffles=0

Merged Map outputs=0

GC time elapsed (ms)=83

CPU time spent (ms)=1270

Physical memory (bytes) snapshot=265928704

Virtual memory (bytes) snapshot=2593497088

Total committed heap usage (bytes)=454033408

Peak Map Physical memory (bytes)=265928704

Peak Map Virtual memory (bytes)=2593497088

File Input Format Counters

Bytes Read=0

File Output Format Counters

Bytes Written=14985

23/05/21 13:57:18 INFO mapreduce.ImportJobBase: Transferred 14.6338 KB in 17.1419 seconds (874.1722 bytes/sec)

23/05/21 13:57:18 INFO mapreduce.ImportJobBase: Retrieved 274 records.

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

9.update “specialty cheeses” to “specialtycheese” and transfer only updated

rows in the above-created directory.

mysql> alter table product\_aisles add column time\_stamp timestamp;

Query OK, 0 rows affected (0.51 sec)

Records: 0 Duplicates: 0 Warnings: 0

mysql> alter table product\_aisles add specialtycheese varchar(55);

Query OK, 0 rows affected (0.56 sec)

Records: 0 Duplicates: 0 Warnings: 0

mysql> update product\_aisles set aisle=specialtycheese where aisle="specialty cheeses";

Query OK, 274 rows affected (0.05 sec)

Rows matched: 274 Changed: 274 Warnings: 0

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

[raghavendr48edu@ip-10-1-1-204 ~]$ sqoop import --connect jdbc:mysql://ip-10-1-1-204.ap-south-1.compute.internal/raghavendr48edu --username raghavendr48edu --password RedKangaroo99$ --table product\_aisles --in

cremental append --check-column specialtycheese --target-dir /user/raghavendr48edu/specialtycheese/ --driver com.mysql.jdbc.Driver -m 1

Warning: /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.

Please set $ACCUMULO\_HOME to the root of your Accumulo installation.

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See  <http://www.slf4j.org/codes.html#multiple_bindings> for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

23/05/21 14:09:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7-cdh6.2.1

23/05/21 14:09:46 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.

23/05/21 14:09:46 WARN sqoop.ConnFactory: Parameter --driver is set to an explicit driver however appropriate connection manager is not being set (via --connection-manager). Sqoop is going to fall back to org.

apache.sqoop.manager.GenericJdbcManager. Please specify explicitly which connection manager should be used next time.

23/05/21 14:09:46 INFO manager.SqlManager: Using default fetchSize of 1000

23/05/21 14:09:46 INFO tool.CodeGenTool: Beginning code generation

23/05/21 14:09:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 14:09:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 14:09:47 INFO orm.CompilationManager: HADOOP\_MAPRED\_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce

23/05/21 14:09:48 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-raghavendr48edu/compile/cac8962fdf0be1bc3e6a1da35b4162ac/product\_aisles.java to /home/raghavendr48edu/./product\_aisles.java. Error: D

estination '/home/raghavendr48edu/./product\_aisles.java' already exists

23/05/21 14:09:48 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-raghavendr48edu/compile/cac8962fdf0be1bc3e6a1da35b4162ac/product\_aisles.jar

23/05/21 14:09:48 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(specialtycheese) FROM product\_aisles

23/05/21 14:09:48 ERROR tool.ImportTool: **Import failed: Character column (specialtycheese) can not be used to determine which rows to incrementally import.**

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

[raghavendr48edu@ip-10-1-1-204 ~]$ **lastvalue=$(mysql -u raghavendr48edu -p -h ip-10-1-1-204.ap-south-1.compute.internal -D raghavendr48edu -e "SELECT max(time\_stamp) from product\_aisles" -sN)**

Enter password:

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*

[raghavendr48edu@ip-10-1-1-204 ~]$ **sqoop import --connect jdbc:mysql://ip-10-1-1-204.ap-south-1.compute.internal/raghavendr48edu --username raghavendr48edu --password RedKangaroo99$ --table product\_aisles --i**

**ncremental lastmodified --check-column time\_stamp --last-value "$lastvalue" --target-dir /user/raghavendr48edu/product\_aisles\_updated/ --driver com.mysql.jdbc.Driver -m 1**

Warning: /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.

Please set $ACCUMULO\_HOME to the root of your Accumulo installation.

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See  <http://www.slf4j.org/codes.html#multiple_bindings> for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

23/05/21 14:16:13 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7-cdh6.2.1

23/05/21 14:16:13 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.

23/05/21 14:16:13 WARN sqoop.ConnFactory: Parameter --driver is set to an explicit driver however appropriate connection manager is not being set (via --connection-manager). Sqoop is going to fall back to org.

apache.sqoop.manager.GenericJdbcManager. Please specify explicitly which connection manager should be used next time.

23/05/21 14:16:13 INFO manager.SqlManager: Using default fetchSize of 1000

23/05/21 14:16:13 INFO tool.CodeGenTool: Beginning code generation

23/05/21 14:16:14 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 14:16:14 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 14:16:14 INFO orm.CompilationManager: HADOOP\_MAPRED\_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce

23/05/21 14:16:16 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-raghavendr48edu/compile/5f030e7366d987ccd94e7a4b050a7529/product\_aisles.java to /home/raghavendr48edu/./product\_aisles.java. Error: D

estination '/home/raghavendr48edu/./product\_aisles.java' already exists

23/05/21 14:16:16 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-raghavendr48edu/compile/5f030e7366d987ccd94e7a4b050a7529/product\_aisles.jar

23/05/21 14:16:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 14:16:16 INFO tool.ImportTool: Incremental import based on column time\_stamp

23/05/21 14:16:16 INFO tool.ImportTool: Lower bound value: '2023-05-21 14:07:37'

23/05/21 14:16:16 INFO tool.ImportTool: Upper bound value: '2023-05-21 14:16:16.0'

23/05/21 14:16:16 INFO mapreduce.ImportJobBase: Beginning import of product\_aisles

23/05/21 14:16:17 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar

23/05/21 14:16:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM product\_aisles AS t WHERE 1=0

23/05/21 14:16:17 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps

23/05/21 14:16:17 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

23/05/21 14:16:17 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/raghavendr48edu/.staging/job\_1684298513961\_0946

23/05/21 14:16:19 INFO db.DBInputFormat: Using read commited transaction isolation

23/05/21 14:16:19 INFO mapreduce.JobSubmitter: number of splits:1

23/05/21 14:16:19 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled

23/05/21 14:16:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1684298513961\_0946

23/05/21 14:16:19 INFO mapreduce.JobSubmitter: Executing with tokens: []

23/05/21 14:16:19 INFO conf.Configuration: resource-types.xml not found

23/05/21 14:16:19 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

23/05/21 14:16:19 INFO impl.YarnClientImpl: Submitted application application\_1684298513961\_0946

23/05/21 14:16:19 INFO mapreduce.Job: The url to track the job:  <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684298513961_0946/>

23/05/21 14:16:19 INFO mapreduce.Job: Running job: job\_1684298513961\_0946

23/05/21 14:16:30 INFO mapreduce.Job: Job job\_1684298513961\_0946 running in uber mode : false

23/05/21 14:16:30 INFO mapreduce.Job: map 0% reduce 0%

23/05/21 14:16:41 INFO mapreduce.Job: map 100% reduce 0%

23/05/21 14:16:41 INFO mapreduce.Job: Job job\_1684298513961\_0946 completed successfully

23/05/21 14:16:41 INFO mapreduce.Job: Counters: 33

File System Counters

FILE: Number of bytes read=0

FILE: Number of bytes written=247336

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=85

HDFS: Number of bytes written=18821

HDFS: Number of read operations=6

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

HDFS: Number of bytes read erasure-coded=0

Job Counters

Launched map tasks=1

Other local map tasks=1

Total time spent by all maps in occupied slots (ms)=7425

Total time spent by all reduces in occupied slots (ms)=0

Total time spent by all map tasks (ms)=7425

Total vcore-milliseconds taken by all map tasks=7425

Total megabyte-milliseconds taken by all map tasks=7603200

Map-Reduce Framework

Map input records=274

Map output records=274

Input split bytes=85

Spilled Records=0

Failed Shuffles=0

Merged Map outputs=0

GC time elapsed (ms)=138

CPU time spent (ms)=2140

Physical memory (bytes) snapshot=274804736

Virtual memory (bytes) snapshot=2594250752

Total committed heap usage (bytes)=469762048

Peak Map Physical memory (bytes)=274804736

Peak Map Virtual memory (bytes)=2594250752

File Input Format Counters

Bytes Read=0

File Output Format Counters

Bytes Written=18821

23/05/21 14:16:41 INFO mapreduce.ImportJobBase: Transferred 18.3799 KB in 24.4268 seconds (770.5076 bytes/sec)

23/05/21 14:16:41 INFO mapreduce.ImportJobBase: Retrieved 274 records.

23/05/21 14:16:41 INFO tool.ImportTool: Final destination exists, will run merge job.

23/05/21 14:16:41 INFO tool.ImportTool: Moving data from temporary directory \_sqoop/f9c8d92be86e47e2aab1b99a73e54a35\_product\_aisles to final destination /user/raghavendr48edu/product\_aisles\_updated

23/05/21 14:16:41 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments:

23/05/21 14:16:41 INFO tool.ImportTool: --incremental lastmodified

23/05/21 14:16:41 INFO tool.ImportTool: --check-column time\_stamp

23/05/21 14:16:41 INFO tool.ImportTool: --last-value 2023-05-21 14:16:16.0

23/05/21 14:16:41 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')

[raghavendr48edu@ip-10-1-1-204 ~]$

*From <*[*https://bigdata.cloudlabs.edureka.co/shell/*](https://bigdata.cloudlabs.edureka.co/shell/)*>*